Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

# An Active Learning Approach to Coreference Resolution

**Mrinmaya Sachan, Eduard Hovy, Eric P. Xing**
School of Computer Science
Carnegie Mellon University
{mrinmays, hovy, epxing}@cs.cmu.edu

## Abstract

In this paper, we define the problem of coreference resolution in text as one of clustering with pairwise constraints where human experts are asked to provide pairwise constraints (pairwise judgments of coreferentiality) to guide the clustering process. Positing that these pairwise judgments are easy to obtain from humans given the right context, we show that with significantly lower number of pairwise judgments and feature-engineering effort, we can achieve competitive coreference performance. Further, we describe an active learning strategy that minimizes the overall number of such pairwise judgments needed by asking the most informative questions to human experts at each step of coreference resolution. We evaluate this hypothesis and our algorithms on both entity and event coreference tasks and on two languages.

## 1 Introduction

Coreference resolution in text is the process of determining when two mentions (named, nominal or pronominal entity mentions, event mentions, etc.) refer to the same identity in the real world. Coreference is a fundamental problem in NLP, an important step in achieving a deeper understanding of the text and is potentially useful for many downstream applications such as paraphrase detection, textual entailment, summarization, question answering, etc.

Despite significant interest over many years with formal evaluation tasks and standardized datasets (MUC, ACE NIST, CONLL shared task, etc.), the problem of coreference remains an unsolved one. Moreover, the applicability and accuracy of prominent coreference systems on real world web applications (often multilingual and noisy) continues to be low. This is because:

1) Accuracies of unsupervised coreference systems is low. On the other hand, obtaining a full coreference annotation is difficult and time consuming. Annotators typically have to make a large number of global decisions, painstakingly seeking answers to difficult questions like the optimal number of coreference chains, the best chain a mention should be assigned to, etc. while respecting transitivity constraints, etc. This limits their applicability to new domains and languages.

2) Coreference is intricately governed by short and long distance semantic constraints [Haghighi and Klein, 2010] and measuring semantic compatibility continues to be an "uphill battle" [Durrett and Klein, 2013].

3) Coreference systems are typically trained on noise-free text and their performance drops considerably on noisy text such as social media.

4) Building a coreference system requires significant feature (or rule) engineering making it difficult to adapt them to other domains, other languages, etc. This again limits their utility in growing body of text data on the web such as newswire, blogs and social media.

However, humans can much more easily wade through the complexities of semantics and noisy data and determine coreferentiality of two mentions often by just looking at the mentions and some additional context when required. Hence, our learning algorithms can gain significantly if they are allowed the assistance of expert humans to answer some of the more difficult questions.

In this paper, we describe a modular, easy-to-use and generalizable technique for coreference. The technique poses the problem of coreference as one of clustering over a feature space of multiple local similarities on two mentions. Interestingly, this reduces the need for heavy weight feature/rule engineering needed for coreference and allows us to have a general formulation for coreference and encapsulates nominal, pronominal entity coreference, event coreference, etc. Using this lightweight technique, we also describe an active learning strategy for coreference. The questions posed by our learner are binary queries about coreferentiality of two mentions [1]. The binary human judgments are modeled as additional constraints on the clustering. Since such manual assistance is expensive, we also let our learner choose the most informative questions for humans given a snapshot of the clustering. The number of judgments required by our learner to learn a competitive coreference system is lower than the number required in a normal supervised setting in terms of the annotation cost (This improvement is more pronounced in noisy data scenarios). Our lightweight technique makes it easier to build a coreference resolution system for new languages and new

[1] We posit that it is easier for humans to provide such pairwise judgments than providing direct judgments about the number of coreference chains, the coreference chain a mention belongs to, etc. Our annotation study on the Hindi coreference dataset supports this.

domains. Our solution is generalizable to many definitions of coreference. We illustrate this by building a competitive entity coreference system in English and several state-of-the-art coreference systems: an event coreference system in English, an entity coreference system in Hindi and an entity coreference resolution system in English for noisy blogs.

## 2 Related Work

Coreference has been a very actively researched area in NLP. These techniques can be broadly categorized into pairwise and global approaches [Ma *et al.*, 2014]. Pairwise approaches [Bengtson and Roth, 2008; Finkel and Manning, 2008] decide if two mentions are coreferent or not using a single function over a diverse set of lexical, syntactic, semantic, and discourse level features derived using the information local to the two mentions. After building a pairwise model to judge coreferentiality or not, these judgments are aggregated to obtain final coreference chains via some heuristics [Stoyanov *et al.*, 2009] or yet another model [Finkel and Manning, 2008]. In contrast to these methods, global techniques [Haghighi and Klein, 2010; Fernandes *et al.*, 2012; Durrett and Klein, 2013] instead of making pairwise local decisions leverage the global structure of the problem. We directly model the process of coreference resolution as a global model that performs clustering over a rich feature space. However, we would like to point out that as shown in our annotation study, it is much easier for humans to provide such pairwise judgments than providing direct judgments about the number of coreference chains, the coreference chain a mention belongs to, etc. Hence we build a global model that can account for global structure and yet allow for pairwise annotation by modeling our problem as constrained clustering.

Metric learning given constraints and constrained pairwise clustering have both been well-studied in machine learning [Cohn *et al.*, 2003; Xing *et al.*, 2003]. Here, most related to our work is [Bilenko *et al.*, 2004] which proposed a technique for integrating constraints and metric learning in semi-supervised clustering and [Basu *et al.*, 2004a] which proposed a technique for actively selecting informative pairwise constraints to get improved clustering performance within a pairwise clustering framework. There is a large body of work in this direction and we cannot completely survey all of them in this article. However, we should point out that our work also has connections with constrained Spectral Clustering [Kamvar *et al.*, 2003], Constrained K-means [Wagstaff *et al.*, 2001], Semi-supervised clustering [Kulis *et al.*, 2005]. In principle, all these approaches can be adapted to our problem.

While active learning, and in particular, active learning for clustering has been hotly researched, surprisingly, despite the obvious importance of the problem, there has not been enough research on active learning for coreference resolution. To the best of our knowledge, the only works on active learning for coreference resolution have been [Gasperin, 2009], [Laws *et al.*, 2012] and [Miller *et al.*, 2012]. While [Gasperin, 2009] reports negative results (not better than random sampling) on a mention-pair model on a biomedical corpus using uncertainty sampling, [Laws *et al.*, 2012] and [Miller *et al.*, 2012] are restricted to active document selection to reduce coreference annotation effort. On the other end, we work on the problem of actively selecting mention pairs, which is a much harder problem. As our approach works on mention-pair level, it can help reduce annotation effort over and above the document-selection model.

We also built a coreference resolution system in Hindi. There has been very little previous work on coreference resolution in Hindi. To the best of our knowledge, the only notable works for coreference resolution in Hindi are [Dakwale *et al.*, 2013], [Dakwale *et al.*, 2012] and [Dutta *et al.*, 2008].With the exception of [Dakwale *et al.*, 2013], the others are rule-based. We compared our method to [Dakwale *et al.*, 2013] and showed superior performance in our experiments. Also, to the best of our knowledge, there is little work that looks at coreference resolution on blogs.

## 3 Method

**Cross-Document Coreference:** We first describe the setup for cross-document coreference resolution. Given a document set $\mathbf{D}$ containing a set of (event/entity) mentions $\mathbf{M}$, a positive judgment of coreferentiality of two mentions (say, $m_i, m_j \in \mathbf{M}$) implies a must-link constraint which penalizes the model with a penalty of $P_{ml}(m_i, m_j)$ if the two mentions are not assigned to the same cluster. Similarly, a negative judgment of coreferentiality of two mentions implies a cannot link constraint which penalizes the model with a weight $P_{cl}(m_i, m_j)$ if the model assigns the entities in the same cluster. Let $\mathbf{C}$ be a clustering over mentions and the number of clusters be $k$. Let $l_i$ denote the cluster-membership indicator that maps each mention $m_i \in M$ to one of the clusters. Also, let $\mathbf{f}(m_i, m_j) \in \mathbb{R}^d$ be the feature representation of a given pair of mentions $m_i, m_j \in \mathbf{M}$ (Features encode various similarities between mentions). Let $\mathbf{ML}$ be the set of must link constraints and $\mathbf{CL}$ be the set of cannot link constraints, respectively. Also let $\mu_c$ be the representation of the medoid of the $c^{th}$ cluster. Let $\mathbf{a} \in \mathbb{R}^d$ be a real-valued vector which defines the weight over local pairwise similarities. $\mathbf{a}$ can be seen as a metric as it determines the overall similarity between points given a host of local similarity features. However, we note that it is technically not a metric as it is not constrained to be positive or follow the triangle inequality.

Here, we note that our approach is inspired from two previously proposed approaches for clustering with pairwise constraints: [Basu *et al.*, 2004a] and [Basu *et al.*, 2004b]. However, both approaches cannot be directly adapted for coreference resolution. This is because both these approaches assume that we have a representation of all mentions (points). While we use the word “clustering” to motivate and explain our solution, we note that we do not have a good representation of mentions (points) as assumed in these works. However, we can effectively define a large number of linguistic features (features encode various views of similarities between the mentions) for each pair of mentions that are very effective for this task. To tackle this, we introduced a feature function for each pair of mentions and slightly change our objective. In our experiments, we will introduce a baseline which are closely related to [Basu *et al.*, 2004a] and [Basu *et al.*, 2004b] and we will show that our approach outperforms

$$a)\quad \mathcal{J} = \sum_{m_i \in \mathbf{M}} \sum_{c=1}^{k} \Big( (-1)^{\delta(l_i \neq c)}\ \mathbf{a^T f}(m_i, \mu_c) \Big) \quad - \sum_{\substack{(m_i,m_j)\in \mathbf{ML} \\ l_i \neq l_j}} P_{ml}(m_i, m_j) \quad - \sum_{\substack{(m_i,m_j)\in \mathbf{CL} \\ l_i = l_j}} P_{cl}(m_i, m_j) \quad - \lambda ||\mathbf{a}||^2$$

$$b)\quad \mathcal{J}_d = \sum_{m_i \in \mathbf{M_d}} \sum_{c=1}^{k_d} \Big( (-1)^{\delta(l_i \neq c)} \mathbf{a^T f}(m_i, \mu_c) \Big) \quad - \sum_{\substack{(m_i,m_j)\in \mathbf{ML_d} \\ l_i \neq l_j}} P_{ml}(m_i, m_j) \quad - \sum_{\substack{(m_i,m_j)\in \mathbf{CL_d} \\ l_i = l_j}} P_{cl}(m_i, m_j)$$

$$c)\quad \mathcal{J} = \sum_{m_i \in \mathbf{M}} \mathbf{a^T f}(m_i, \mu_{c_i}) \quad - \sum_{(m_i,m_j)\in \mathbf{ML} l_i \neq l_j} P_{ml}(m_i, m_j) \quad - \sum_{\substack{(m_i,m_j)\in \mathbf{CL} \\ l_i = l_j}} P_{cl}(m_i, m_j) \quad - \lambda ||\mathbf{a}||^2$$

Figure 1: The coreference objectives: a) cross-document coreference objective, b) per-document in-document coreference objective and c) cross-document BL baseline

them in practice.

Our solution is to combine the notion of pairwise constraints and clustering through the objective $\mathcal{J}$ in Figure 1a. Note that the penalties ($P_{ml}(m_i, m_j)$ and $P_{cl}(m_i, m_j)$ for violating the must-link and cannot-link constraints between entities $m_i$ and $m_j$, respectively) are fairly general. Intuitively, one can say that the penalty for violating a must-link constraint between distant points should be higher than that between nearby points. This captures the fact that if two must-linked points are far apart according to the current metric, the metric is grossly inadequate and needs severe modification. Analogously, the penalty for violating a cannot-link constraint between two points that are nearby according to the current metric should be higher than for two distant points. To reflect this intuition, we set $P_{ml}(m_i, m_j) = -w_{ml} \times \mathbf{a^T f}(m_i, m_j)$ and $P_{cl}(m_i, m_j) = w_{cl} \times \mathbf{a^T f}(m_i, m_j)$. The problem (maximize $\mathcal{J}$) can be solved using a variation of the familiar hard-EM solution described in Algorithm 1 for k-medoids where we also update the metric in the M-step.

**The HMRF Interpretation:** Like many classic previous works in constrained k-means [Basu *et al.*, 2004a; 2004b], it can be shown that the objectives can be written as the negative logarithm of the posterior probability (proportional to the configuration energy) of a Hidden Markov Random Field (HMRF) over the data with a well-defined potential function and noise model. Let the metric $\mathbf{a}$ be drawn under the model $\mathbf{a} \sim \mathcal{N}(0, \frac{1}{2\lambda})$ and the cluster assignments $l$ be hidden. Consider the MRF defined over $\mathbf{M}$ where the the field (set of random variables) $\boldsymbol{\mathcal{F}} = \{F_m\}_{m=1}^{\mathbf{M}}$ is such that each random variable $F_m$ can take a value equal to $l_m$ in $\{1, \ldots, k\}$. Let the clique potentials be defined between pairs of points. By Hammersley-Clifford theorem, probability of configuration $\mathcal{L} = \{l_m\}_{m=1}^{k}$ is given by the gibbs distribution $P(\mathcal{L}) = \frac{1}{Z} exp\left(-V(\mathcal{L})\right) = \frac{1}{Z} exp\left(\sum_{i,j} V_{(i,j)}(l_i, l_j)\right)$.

$$V_{(i,j)}(l_i, l_j) = \begin{cases} P_{ml}(m_i, m_j) & \text{if } (m_i, m_j) \in \mathbf{ML} \\ P_{cl}(m_i, m_j) & \text{if } (m_i, m_j) \in \mathbf{CL} \\ 0 & \text{otherwise} \end{cases}$$

Furthermore, we assume that given $\mathbf{a}$ and $\{\mu_c\}_{c=1}^{k}$, the data points are generated i.i.d. via the noise model:

$$P(\mathbf{M}|\mathcal{L}) \propto exp\left( \sum_{m_i \in \mathbf{M}} \sum_{c=1}^{k} \Big( (-1)^{\delta(l_i \neq c)}\ \ \mathbf{a^T f}(m_i, \mu_c) \Big) \right)$$

. It is easy to see that MAP estimation of the HMRF is equivalent to maximizing $\mathcal{J}$.

**In-Document Coreference:** The above technique can easily be extended to in-doc coreference. Given a document set $\mathbf{D}$ where each document $d \in \mathbf{D}$ contains a set of (event/entity) mentions $\mathbf{M_d}$ and a bunch of must link and cannot-link constraints ($\mathbf{ML_d}$, $\mathbf{CL_d}$), to be clustered into $k_d$ clusters. The problem is similarly posed as one of maximizing $\mathcal{J}$ ($\mathcal{J}$ is defined in terms of per-document objectives $\mathcal{J}_d$ shown in Figure 1b) as $\mathcal{J} = \sum\limits_{d \in \mathbf{D}} \mathcal{J}_d - \lambda ||\mathbf{a}||^2$.

**High Precision Singleton removal:** Empirically, we noticed that a large number (about 70 percent) of coreference chains (clusters) in our datasets were singletons. Singletons could potentially obfuscate the clustering process. Hence, we employed a high precision singleton detector that identifies and removes some singletons before feeding to our clustering algorithm. For each mention, the sieve computes the similarity of the mention with all other candidate mentions (with $\mathbf{a} = \mathbf{1}$) and removes it if the ratio of the maximum similarity between the mention and all other candidate mentions and the average similarity between all pairs of candidate mentions is less than a threshold (tuned on dev set).

**Choosing the number of clusters:** Choosing the right number of clusters ($k^*$) is an important step and has a significant impact on final accuracies. We employ the Bayesian Information Criterion [Schwarz, 1978]: $k^* = \arg\max\limits_{k} [\mathcal{J}(k) - \lambda' \ \textit{log(k)}]$ for cross-doc coreference and $k_d^* = \arg\max\limits_{k} [\mathcal{J}_d(k) - \lambda' \ \textit{log(k)}]$ for in-doc coreference.

**Initialization** It is well known that the choice of initialization is important in EM style solutions for K-means because of several local optima. This holds for our technique too. We resort to multiple initializations and pick the solution with highest objective. The initialization step is fairly simple: Given a set of must-link and cannot-link constraints, we compute the transitive closure ($ML(m_i, m_j) \wedge ML(m_j, m_k) \implies ML(m_i, m_k)$ and $CL(m_i, m_j) \wedge ML(m_i, m_k) \implies CL(m_j, m_k)$). Then, given these expanded set of constraints, we initialize ($\mathbf{C}$ and $\mu$). We initialize the metric $\mathbf{a}$ as a vector of all 1's (initial similarity is thereby set to the sum of all local feature similarities).

**Feature Design:** As described before, our technique is generalizable to both entity and event coreference. Our fea-

**Algorithm 1** Cross-Document CorefSolver($\mathbf{M}$, $\mathbf{ML}$, $\mathbf{CL}$)

```
Initialize Metric, Random Clustering, Cluster Medoids
while Not converged do:
    E-step: Reassign points to nearest clusters:                                                     (1)
    $c^*_{m_i} = \arg\min_{c} \left[ \mathbf{a^T f}(m_i, \mu_c) + w_{ml} \sum_{\substack{(m_i,m_j)\in\mathbf{ML} \\ l_i \neq l_j}} \mathbf{a^T f}(m_i, m_j) - w_{cl} \sum_{\substack{(m_i,m_j)\in\mathbf{CL} \\ l_i = l_j}} \mathbf{a^T f}(m_i, m_j) \right]$
                                                                                                     $\forall m_i \in \mathbf{M}$
    M-step:
    (i) Redesignate cluster medoids:                                                                 (2)
    $\mu^*_c = \arg\min_{\mu_c \in \mathbf{M_c}} \sum_{m_i \in \mathbf{M_c}} \mathbf{a^T f}(m_i, \mu_c)$                                   $\forall c \in \mathbf{1} \ldots \mathbf{k}$
    (ii) Update the metric ($\frac{\partial \mathcal{J}}{\partial \mathbf{a}} = 0$):                                                       (3)
    $\mathbf{a} = \frac{1}{\lambda} \left[ \sum_{m_i \in \mathbf{M}} \sum_{c=1}^{k} \mathbf{f}(m_i, \mu_c) + w_{ml} \sum_{\substack{(m_i,m_j)\in\mathbf{ML} \\ l_i \neq l_j}} \mathbf{f}(m_i, m_j) - w_{cl} \sum_{\substack{(m_i,m_j)\in\mathbf{CL} \\ l_i = l_j}} \mathbf{f}(m_i, m_j) \right]$
end while
```

tures essentially encode a bunch of similarities between pairs of mentions. For entity coreference in English, we use the pairwise features employed in the Berkeley Coreference System [Durrett *et al.*, 2013]. The features are a bunch of anaphoricity features, features that encode configurational similarity, match features, local agreement and discourse level features. For event coreference resolution, we use the features in [Liu *et al.*, 2014]. These features also encode various facets of similarities between event mentions: similarities of event lemmas, arguments and predicates, expanded synonym clouds, number, animacy, gender, NE label match, etc. For entity coreference in Hindi, we (quickly) build a small set of local, pairwise similarity features ourselves. Most coreference systems encode a much more complicated set of features (entity level features, features that model repeated chains of pronouns, etc.). However, a small set of local similarity based features suffice in semi-supervised, noisy or low resource language settings. We demonstrate this by quickly building a state-of-the-art coreference system for Hindi and blogs in English. Our features for these two experiments are listed in the supplementary.

**Choosing the Right Questions:** Next, given our clustering model, we explore techniques for actively selecting the most informative constraint. The informativeness of a constraint depends on the current state of the clustering $\mathcal{C}$ (positions of all the mention points dictated by the metric and current cluster assignments). Hence, the overall problem essentially is of selecting the optimal mention pair $(m_i, m_j)^*$ among all pairs. We explore various heuristics :

**1) Uncertainty Model (UM):** Pick the mention pair with the highest uncertainty (entropy) $H$ given the clustering $\mathcal{C}$.

$$(m_i, m_j)^* \quad = \quad \arg\max_{(m_i,m_j)} H(\Omega_{ml} | m_i, m_j, \mathcal{C})$$

**2) Expected Judgement Error (EJE):** Select the mention pair that maximize the expected judgement error:

$$(m_i, m_j)^* = \arg\max_{(m_i,m_j)} \mathbb{E}((\hat{\Omega} - \Omega)^2 | m_i, m_j)$$

$$\mathbb{E}((\hat{\Omega} - \Omega)^2 | m_i, m_j) = (\hat{\Omega} - 1)^2 P(\Omega_{ml} | m_i, m_j) \\ +(\hat{\Omega} + 1)^2 P(\Omega_{cl} | m_i, m_j)$$

**3) Change in Objective (CiO):** Choose the mention pair that causes largest increase in the objective $\mathcal{J}$

**4) Closest Points assigned to different clusters (CP-DC):** Mention pairs closest to each other (according to the metric **a**) which are assigned to different clusters could be good candidates. They may actually be coreferent.

**5) Farthest Points assigned to same cluster (FP-SC):** Mention pairs farthest from each other and assigned to the same cluster may not be coreferent.

**6) Explore and Exploit (E&E):** We employ the explore and exploit strategy for active constraint selection [Basu *et al.*, 2004a]. The algorithm is implemented in two independent steps. The first step, called the *Explore step* is based on a *farthest-first traversal scheme*, which finds $t$ points such that they are far from each other. This is done to explore the data to get $t$ disjoint neighborhoods, each belonging to a different cluster in the underlying clustering of the data. Then, given these disjoint regions, in the *Consolidate step*, the algorithm asks $t - 1$ queries by pairing a randomly selected point with another point in the $t - 1$ distinct neighborhoods to find out the neighborhood to which it belongs.

In the above heuristics, the probabilities can be intuitively estimated using the distances according to the current metric. In particular, we posit $P(\Omega_{ml} | m_i, m_j) = \frac{\mathbf{a}^T \mathbf{f}(m_i, m_j)}{\mathbf{a}^T \mathbf{f}(m', m'')}$ where $m'$ and $m''$ are points that are maximally apart according to the current metric. $P(\Omega_{cl} | m_i, m_j) = 1 - P(\Omega_{ml} | m_i, m_j)$. Apart from the above heuristics, we also employ an **Ensemble** method for constraint selection. The ensemble computes the ratio of the score of the best constraint and the average score over the $\mathbf{M}^2$ constraints for each of the aforementioned six heuristics and picks the constraint with the highest ratio.

Our overall active learning solution is to start with a small set of constraints (initialization) and then iteratively compute the clustering, asking for more supervision intermittently.

# 4 Experiments

**Datasets:** We use four datasets for our evaluation. First is the standard cross-document coreference evaluation dataset in English, ACE-2008 [Strassel *et al.*, 2008]. We use the berkeley coreference system [Durrett *et al.*, 2013] for mention detection which in turn derives its mentions from the Stanford system [Lee *et al.*, 2011]. Next, we will also use the (in-document) IC Event Coreference Corpus and setup [Liu *et al.*, 2014] for evaluations on event coreference and a small newswire dataset on Hindi annotated for in-document entity coreference by us. To create the Hindi entity coreference dataset, we annotated 91 news-articles (50 train, 15 dev and 26 test) published on October 10, 2004 in two popular Hindi newspapers *Amar Ujala* and *Navbharat Times* using Brat [Stenetorp *et al.*, 2012]. Time for each annotation was also recorded for comparison with our active learner. Here the gold mentions were provided to our method as well as the baselines. As standard practice, we report the CONLL score. Finally, we also created a small annotated dataset of 100 blogs (in-turn drawn from a very large collection of blogs posts[2]). Here, the 100 documents (50 train, 25 dev, 25 test) were annotated for in-doc coreference resolution.

**Baselines:** To compare against other traditional coreference techniques that take full annotated documents as input: Berkeley [Durrett *et al.*, 2013] and UIUC systems [Bengtson and Roth, 2008], [Dakwale *et al.*, 2013] and [Liu *et al.*, 2014], for a given amount of pairwise annotation $A$, we train them on a random selection of $\lceil D \rceil$ documents, where $D$ documents contain $A$ pairwise annotations. Also, as described before, we construct a baseline (we call it "BL") that is again inspired from our approach and based on [Basu *et al.*, 2004a]. It models cross-document coreference through the objective $\mathcal{J}$ in Figure 1c. The only difference between our technique and "BL" is in the noise model. Our technique minimizes the distances between points in same coreference chain as well as maximizes the distances between points that are in the different coreference chains. However, "BL" only minimizes the distance between points in same coreference chain.

**Evaluation:** To evaluate and compare our technique, we posit an oracle that gives us the must-links and cannot-link constraints (gold annotations from the dataset) on demand (with the exception of the Hindi entity coreference task where we also perform a user study that compares various algorithms in terms of the time spent in providing pairwise or cluster level annotation to them). For each experiment, we run the algorithm 10 times and use the best solution (solution with the highest objective value). For each run of the algorithm, we assume convergence when the change in objective over successive iterations drops below 0.01 percent. The hyper parameters $\lambda$, $\lambda'$, $w_{ml}$ and $w_{cl}$ and the threshold for singleton removal are optimized on the development sets using line search.

First, we investigate our active constraint selection procedures. Figure 2a plots the performance (CONLL F1 score) achieved by our technique against the amount of supervision (as proportion of the training set) provided to the system on the English entity coreference ACE 2008 dataset. First, we can notice that by employing any of the five active constraint selection strategies, we can achieve a higher performance than a random selector. Also, we can observe that the ensemble procedure consistently achieves a better performance than the five selection strategies. Hence, we employ the ensemble procedure in all our future experiments.

Figure 2b plots the CONLL score against the amount of supervision and compares our system to the Berkeley and UIUC coreference systems. We cannot compare against the Stanford system as it is rule-based. Here, our system is worse than the Berkeley system when all the annotations are used (we are about 3 CONLL points behind the Berkeley and Stanford systems). More importantly, it can be seen that our method works much better when amount of annotation is lower. In fact, we continue to achieve better performance up until the stage when the amount of supervision is 70% of the training set. Moreover, we achieve a very competitive coreference performance ($>$ 50 CONLL points) with just 40% annotation. Additionally, we simulate our performance when the human supervision is not perfect. We perform two experiments where our oracle gives judgments with 10% and 20% error rates respectively. This is important as in real world settings, human judgments may not be perfect. This shows that our technique can be practically and effectively used to achieve coreference actively at a low annotator cost. Also, as expected, our method (even with 10% error rate) outperforms the baseline BL.

Since our coreference system models features in the form of local mention similarities, it is fairly general and can be used for all kinds of coreference. As an example, we do a similar exercise in event coreference resolution (see Figure 2c) where we achieve comparable results with the state-of-the-art baseline [Lee *et al.*, 2011] . All our observations for event coreference hold here too. It is interesting to note that since our technique only needs pairwise similarity features, we employed significantly lesser features (34 features as against 139 employed in [Lee *et al.*, 2011]). We believe that our technique has the potential of significantly reducing the painful feature engineering process employed in coreference. We further back this observation with quickly building a state-of-the-art coreference system for Hindi (see Figure 2d). Here, our method outperforms the state-of-the-art [Dakwale *et al.*, 2013] even with a 10 percent error rate. Again, in both experiments, our method outperforms the baseline BL (with the ensemble active learner). All these observations hold in the Entity Coreference experiments on the blogs dataset (see Figure 2e). The blogs dataset is more noisy and more realistic in terms of the kind of text data seen on the web.

**Timed Annotation Exercise:** We also perform an annotation study to show that our method reduces the time taken to elicit annotation. To do so, we annotated the Hindi entity coreference dataset twice: first by answering all binary pairwise questions of coreferentiality and again by performing a cluster level annotation (i.e. answering what cluster (among clusters instantiated so far) an entity is coreferent to (if any))[3]. The pairwise annotation took 44.2 man-hours and

[2] http://www.icwsm.org/2009/data/

[3] Both annotation processes were separated by 3 months to minimize bias due to previous exposure to the data and a document was

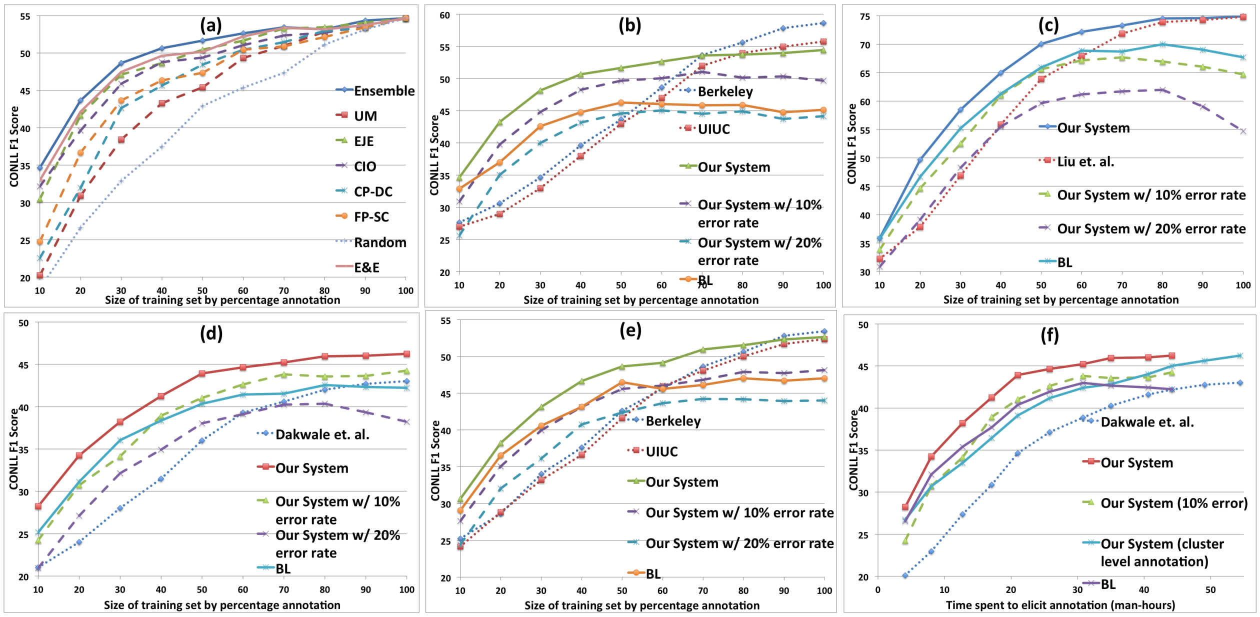

Figure 2: (a) Coreference Performance of various constraint selection methods as size of training data is varied (ACE 2008 dataset), (b, c, d, e) Performance as size of training data is varied on (b) cross-document Entity-Coreference (ACE-2008 dataset), (c) in-document Event-Coreference (IC corpus), (d) in-document Entity-Coreference (Hindi-Newswire), (e) in-document Entity-Coreference (English-Blogs) and (f) Performance vs Time to elicit annotation (in man-hours) for in-document Entity Coreference (Hindi-Newswire)

cluster level annotation took 54.4 man-hours. This supports our hypothesis that humans find it easier to answer pairwise queries as compared to cluster level queries.

Figure 2f plots the performance of our system, a variation of our system where 10% of the pairwise annotation is flipped, a variant of our system where we used cluster-level annotation (by modeling them as pairwise constraints) and [Dakwale *et al.*, 2013] which also uses cluster-level annotation. Interestingly, our method (sometimes even with 10% error rate) works better than the variant of our method which takes cluster level annotation and always better than [Dakwale *et al.*, 2013]. Our method again performs better than "BL". Again, the improvement is more when amount of annotation is low and decreases over time.

**Error Analysis:** We showed that our approach is extremely useful in low-supervision scenarios, making it particularly useful for low-resource languages or text datasets found on the web such as blogs, social media, etc. However, as shown in Figure 2a, we had seen that our approach does not reach the performance of the Berkeley, Stanford and UIUC systems when all the supervision is used. To analyze this, we compared our system against the three systems using the Berkeley Coreference Analyzer [Kummerfeld and Klein, 2013] [4] for the cross-document newswire entity-detection task. Since we use the Berkeley (which in-turn uses the Stanford system) for mention detection, the "span errors" (errors in detecting mentions and their spans) of our system are the same as those in the Berkeley and the Stanford systems. A majority of errors in our approach were of the "merge" and "split" types where the clusters get divided or conflated. We posit that this is because of our model simplification of directly modeling the problem as clustering. Note that clustering has a global objective. The other models (for example, the Berkeley model) often map each mention to its antecedent. This allows them to do better at noun-pronoun coreference resolution with richer features like "it" has a geopolitical entity as its antecedent. Indeed, we observed that a significant proportion of the errors in our system are related to noun-pronoun coreferences, especially when we have a chain of repeated pronouns that refer to the same noun. Notably such phenomena are less prevalent in event coreference. This is perhaps the reason why our system does much better in event coreference - even beating [Liu *et al.*, 2014]. However, we must also note that the global model allows us the simplicity and flexibility and works well even with smaller number of features. Our clustering model also correctly labels some of the cataphora resolutions (when an anaphor precedes its antecedent). The Berkeley system, on the other hand, misses all the cataphora links due to its model design. Our analysis also concurs with [Kummerfeld and Klein, 2013; Durrett and Klein, 2013] in the finding that a majority of the errors in the earlier systems are because of the lack of a good model for semantics. Existing semantic features give only slight benefit because they do not provide strong enough signals for coreference. Our full model also has the same drawback. However, our system makes fewer such errors in its active setting - the human intervention allows the system to solicit supervision for some harder decisions, which require

given to the same annotator both times to minimize annotator bias.

[4] https://code.google.com/p/berkeley-coreference-analyser/

semantic modeling whereas the other systems have no such functionality.

## 5 Conclusion

We began with the observation that well-known coreference systems are trained on *painstakingly-annotated clean* datasets and employ *heavy feature engineering*. This hinders their adaptability to noisy, low resource scenarios commonly encountered in the real world. To mitigate this, we described a technique for solving the general problem of all forms of coreference (nominal and pronominal entity, event coreference, etc.) as one of clustering in a high dimensional feature space with pairwise constraints. The constraints are essentially judgments of coreferentiality for pairs of mentions. Such supervision is much easier to obtain from humans as they can often judge coreferentiality given the right context. Finally, we also presented a set of active learning strategies that can be employed to ask informative questions and obtain a good coreference solution with less annotation.

## References

[Basu *et al.*, 2004a] Sugato Basu, A. Banjeree, ER. Mooney, Arindam Banerjee, and Raymond J. Mooney. Active semi-supervision for pairwise constrained clustering. In *Proceedings of SDM*, 2004.

[Basu *et al.*, 2004b] Sugato Basu, Mikhail Bilenko, and Raymond J Mooney. A probabilistic framework for semi-supervised clustering. In *Proceedings of KDD*, 2004.

[Bengtson and Roth, 2008] Eric Bengtson and Dan Roth. Understanding the value of features for coreference resolution. In *Proceedings of EMNLP*, 2008.

[Bilenko *et al.*, 2004] Mikhail Bilenko, Sugato Basu, and Raymond J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of ICML*, 2004.

[Cohn *et al.*, 2003] David Cohn, Rich Caruana, and Andrew Mccallum. Semi-supervised clustering with user feedback. Technical report, 2003.

[Dakwale *et al.*, 2012] Praveen Dakwale, Himanshu Sharma, and Dipti M Sharma. Anaphora annotation in hindi dependency treebank. In *Proceedings of PACLIC*, 2012.

[Dakwale *et al.*, 2013] Praveen Dakwale, Vandan Mujadia, and Dipti M Sharma. A hybrid approach for anaphora resolution in hindi. In *Proceedings of IJCNLP*, 2013.

[Durrett and Klein, 2013] Greg Durrett and Dan Klein. Easy victories and uphill battles in coreference resolution. In *Proceedings of EMNLP*, 2013.

[Durrett *et al.*, 2013] Greg Durrett, David Hall, and Dan Klein. Decentralized entity-level modeling for coreference resolution. In *Proceedings of ACL*, Sofia, Bulgaria, August 2013.

[Dutta *et al.*, 2008] Kamlesh Dutta, Nupur Prakash, and Saroj Kaushik. Resolving pronominal anaphora in hindi using hobbs algorithm. *Web Journal of Formal Computation and Cognitive Linguistics*, 2008.

[Fernandes *et al.*, 2012] Eraldo Rezende Fernandes, Cícero Nogueira Dos Santos, and Ruy Luiz Milidiú. Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Joint Conference on EMNLP and CoNLL-Shared Task*, 2012.

[Finkel and Manning, 2008] Jenny Rose Finkel and Christopher D. Manning. Enforcing transitivity in coreference resolution. In *Proceedings of ACL-HLT*, 2008.

[Gasperin, 2009] Caroline Gasperin. Active learning for anaphora resolution. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, 2009.

[Haghighi and Klein, 2010] Aria Haghighi and Dan Klein. Coreference resolution in a modular, entity-centered model. In *Proceeding of NAACL*, 2010.

[Kamvar *et al.*, 2003] Kamvar Kamvar, Sepandar Sepandar, Klein Klein, Dan Dan, Manning Manning, and Christopher Christopher. Spectral learning. In *In IJCAI*, 2003.

[Kulis *et al.*, 2005] B. Kulis, S. Basu, I. Dhillon, and Raymond J. Mooney. Semi-supervised graph clustering: A kernel approach. In *Proceedings of ICML*, 2005.

[Kummerfeld and Klein, 2013] Jonathan K. Kummerfeld and Dan Klein. Error-driven analysis of challenges in coreference resolution. In *Proceedings of EMNLP*, 2013.

[Laws *et al.*, 2012] Florian Laws, Florian Heimerl, and Hinrich Schütze. Active learning for coreference resolution. In *Proceedings of NAACL HLT*, 2012.

[Lee *et al.*, 2011] Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the CONLL: Shared Task*, 2011.

[Liu *et al.*, 2014] Zhengzhong Liu, Jun Araki, Eduard Hovy, and Teruko Mitamura. Supervised within-document event coreference using information propagation. In *Proceedings of LREC*, 2014.

[Ma *et al.*, 2014] Chao Ma, Janardhan Rao Doppa, J Walker Orr, Prashanth Mannem, Xiaoli Fern, Tom Dietterich, and Prasad Tadepalli. Prune-and-score: Learning for greedy coreference resolution. In *Proceeding of EMNLP*, 2014.

[Miller *et al.*, 2012] Timothy A. Miller, Dmitriy Dligach, and Guergana K. Savova. Active learning for coreference resolution. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, 2012.

[Schwarz, 1978] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 1978.

[Stenetorp *et al.*, 2012] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. Brat: A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at EACL*, 2012.

[Stoyanov *et al.*, 2009] Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of ACL-IJCNLP*, 2009.

[Strassel *et al.*, 2008] Stephanie Strassel, Mark Przybocki, Kay Peterson, Zhiyi Song, and Kazuaki Maeda. Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction. In *LREC*, 2008.

[Wagstaff *et al.*, 2001] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schroedl. Constrained k-means clustering with background knowledge. In *Proceedings of ICML*, 2001.

[Xing *et al.*, 2003] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. Distance metric learning, with application to clustering with side-information. In *Advances in NIPS*, 2003.