# Domain Adaptation for Coreference Resolution: An Adaptive Ensemble Approach

Jian Bo Yang<sup>†‡\*</sup>, Qi Mao<sup>†</sup>, Qiao Liang Xiang<sup>†</sup>, Ivor W. Tsang<sup>†</sup>, Kian Ming A. Chai<sup>§</sup>, Hai Leong Chieu<sup>§</sup>

<sup>†</sup> School of Computer Engineering, Nanyang Technological University, Singapore
 <sup>‡</sup> Electrical and Computer Engineering Department, Duke University, USA
 <sup>§</sup> DSO National Laboratories, Singapore

### Abstract

We propose an adaptive ensemble method to adapt coreference resolution across domains. This method has three features: (1) it can optimize for any user-specified objective measure; (2) it can make document-specific prediction rather than rely on a fixed base model or a fixed set of base models; (3) it can automatically adjust the active ensemble members during prediction. With simplification, this method can be used in the traditional withindomain case, while still retaining the above features. To the best of our knowledge, this work is the first to both (i) develop a domain adaptation algorithm for the coreference resolution problem and (ii) have the above features as an ensemble method. Empirically, we show the benefits of (i) on the six domains of the ACE 2005 data set in domain adaptation setting, and of (ii) on both the MUC-6 and the ACE 2005 data sets in within-domain setting.

# 1 Introduction

Coreference resolution is a fundamental component of natural language processing (NLP) and has been widely applied in other NLP tasks (Stoyanov et al., 2010). It gathers together noun phrases (mentions) that refer to the same real-world entity (Ng and Cardie, 2002). In the past decade, several coreference resolution systems have been proposed, e.g., (Ng and Cardie, 2002), (Denis and Baldridge, 2007) and (Stoyanov et al., 2010). All of these focus on the within-domain case — to use the labeled documents from a domain to predict on the unlabeled documents in the same domain. However, in practice, there is usually limited labeled data in a specific domain of interest, while there may be plenty of labeled data in other related domains. Effective use of data from the other domains for predicting in the domain of interest is therefore an important strategy in NLP. This is called domain adaptation, and, in this context, the former domains is called the *source domains*, while the latter domain is called the *target domain* (Blitzer et al., 2006; Jiang and Zhai, 2007).

Based on the type of the knowledge to be transferred to the target domain, domain adaptation learning can be categorized as instance-based method, feature-based method, parameter-based method or relational-knowledge-based method (Pan and Yang, 2010). Previously, domain adaptation learning has been successfully used in other NLP tasks such as relation extraction (Jiang, 2009) and POS tagging (Jiang and Zhai, 2007), semantic detection (Tan et al., 2008), name entity recognition (Guo et al., 2009) and entity type classification (Jiang and Zhai, 2007). However, to the best of our knowledge, it has yet to be explored for coreference resolution.

In this paper, we propose an adaptive ensemble method to adapt coreference resolution across domains. This proposed method can be categorized as both feature-based and parameter-based domain adaptation learning methods. It has three main steps: *ensemble creation, cross-domain knowledge learning* and *decision inference*. The first step creates the ensemble by collecting a set of base models, which can be any individual methods with various features/instances/parameters settings. The second step analyzes the collected base models from vari-

<sup>\*</sup>The work is done during postdoc in NTU, Singapore.

ous domains and learns the cross-domain knowledge between each target domain and the source domain. The third step infers the final decision in the target domain based on all ensemble results.

In addition to domain adaptation, the proposed adaptive ensemble method has the following features that are absent in the other ensemble methods. First, it can optimize any *user-specified objective measure* without using a separate development set. Second, it can provide *document-specific prediction* instead of relying on a fixed base model or a fixed set of base models for all documents. Third, it can automatically *adjust the active ensemble members* in decision inference so that underperforming base models are filtered out. The proposed method can also be used in the traditional within-domain problem with some simplifications.

We conduct experiments for coreference resolution under both the within-domain setting and the domain-adaptation setting. In the within-domain setting, we compare the proposed adaptive ensemble method with the mention-pair methods and other ensemble methods on the MUC-6 and ACE 2005 corpora. The results show that the proposed adaptive ensemble method consistently outperforms these baselines. In the domain adaptation setting, we use the ACE 2005 corpora to create six domain adaptation tasks to evaluate the effectiveness of our domain adaptation learning. The results show that our method outperforms baselines that do not use domain adaptation.

The paper is organized as follows. Section 2 reviews some existing ensemble methods for coreference resolution. Section 3 presents the proposed adaptive ensemble method for domain adaptation problems. Section 4 presents a special case of the proposed method for the within-domain setting. Section 5 presents the experiments under both the within-domain and the domain adaptation settings. We conclude and discuss future work in Section 6.

### 2 Existing Ensemble Methods

Many ensemble methods have been proposed in the machine learning literature, e.g., bagging (Breiman, 1996), boosting (Freund and Schapire, 1996), random forest (Breiman, 2001) and mixture models (Bishop, 2007). Some of them have been success-

fully used in coreference resolution (Pang and Fan, 2009; Munson et al., 2005; Rahman and Ng, 2011a). However, these methods only focus on the within-domain setting.

All these methods comprise of two steps: ensemble creation and decision inference. Ng and Cardie (2003) and Vemulapalli et al. (2009) applied the bagging and boosting techniques on the documents to create the ensemble. Recently, Rahman and Ng (2011a) further enriched the ensemble by considering various feature sets and learning models. Specifically, three types of feature sets (conventional, lexical and combined) and three learning algorithms (mention-pair model, mention-ranking model and the clustering-ranking model) are employed. In decision inference, these methods used voting or averaging to get the final prediction. Rahman and Ng (2011a) proposed four voting strategies for prediction: applying best Per-NP-Type model, antecedentbased voting, cluster-based voting and weighted clustering-based voting. Although their approaches achieved promising results in their end-to-end systems, these do not consider the user-specific performance measure during the ensemble learning.

Another branch of ensemble methods uses model selection (Munson et al., 2005; Ng, 2005), similar to the conventional model selection method for generic parameter-tuning. The method of (Munson et al., 2005) first collects a large family of base models. Then, a separate tuning set with ground truth is used to evaluate each base model's performance. Finally, an iterative approach is used to select the best performed base models to form the ensemble. Like other methods, this method uses the average strategy in decision inference. Similarly, the method of (Ng, 2005) ranks base models according to their performance on separate tuning set, and then uses the highest-ranked base model for predicting on test documents. These methods require a separate set of labeled documents to assess the generalization performance.

# 3 Adaptive Ensemble Method

In this section, we give our adaptive ensemble method for domain adaptation for coreference resolution. We first introduce some notations.

For a corpus of N documents, document  $\mathcal{D}_i$ 

is the  $i^{th}$  document, and it contains  $n_i$  mentions  $\mathbf{m}_i = (m_i^1, \ldots, m_i^{n_i})$  with the ordering of each mention as they appear in the document. The index set of all mention pairs in  $\mathcal{D}_i$  is  $\mathcal{E}_i = \{(a, b) \mid 1 \le a < b \le n_i\}$ . The transpose of vector x is x'. The performance measure function for document  $\mathcal{D}$  is  $\Lambda(q(\mathcal{D}); f(\mathcal{D}))$ , where  $q(\mathcal{D})$  and  $f(\mathcal{D})$  represent the coreference ground-truth and prediction by model f on document  $\mathcal{D}$  respectively. In coreference resolution, typical performance measure functions include MUC (Vilain et al., 1995), Rand index (Rand, 1971), B-CUBED (Bagga and Baldwin, 1998) and CEAF (Luo, 2005). In this paper,  $\Lambda$  can either be used as part of an objective function in learning or as an evaluation measure for assessing the performance of a coreference system.

We consider the typical domain adaptation problem, which has one target domain t and p  $(p \ge 1)$ source domains  $s_1, \ldots, s_p$ . The target domain contains  $N^{(t)}$  labeled documents and M unlabeled documents, while source domains contain  $N^{(s_1)}, \ldots, N^{(s_p)}$  labeled documents. Unlabeled data in the source domains are not used. We use  $\mathcal{D}_i^{(v)}$  for the  $i^{th}$  document in domain v.

### 3.1 Ensemble Creation

Mention-pair methods have been widely-used for coreference resolution due to their efficiency and effectiveness, and they have often been taken as base models in ensemble learning (Rahman and Ng, 2011a; Munson et al., 2005). We adopt a similar approach by using the standard mention-pair method (Soon et al., 2001; Ng and Cardie, 2002) with various parameters to form the ensemble, though our framework can incorporate other coreference methods in the ensemble. Mention-pair methods usually comprise of two steps. The first step classifies every mention pair into either coreference or noncoreference with a confidence between 0 and 1. The second step partitions the set of mentions into clusters based on the confidence values, where mentions in each cluster are presumed to be the same underlying entity.

**Classification** We use Soon's approach (Soon et al., 2001) to select a portion of mention pairs to train a binary classifier because this has better generalization (Soon et al., 2001). The positive mention pairs

are the anaphoric mention  $m_i^b$   $(b = 2, ..., n_i)$  paired with its closest antecedent mention  $m_i^a$  (a < b), while the negative mention pairs are the mention  $m_i^b$  paired with each of the intervening mentions  $m_i^{a+1}, m_i^{a+2}, ..., m_i^{b-1}$ . Following (Rahman and Ng, 2011a), our binary classifier is SVM with the regularization parameter C. The classifier is trained with the software Liblinear (Fan et al., 2008), which is also used to give probabilistic binary predictions.

**Clustering** We adopt closest-first clustering (Soon et al., 2001) and best-first clustering (Ng and Cardie, 2002) to determine whether a mention pair is coreferent. For each mention, the closest-first method (or best-first method) links it to the the closest (or the best) preceding mention if the confidence value (obtained from the first step) of this mention pair is above a specified threshold t.

**Features** For each mention pair, we use the d = 39 features proposed by Rahman and Ng (2011b) to represent it. These features can be extracted using the Reconcile software (Stoyanov et al., 2010). We use  $\hat{\phi}_{a,b} \in \mathbb{R}^d$  to represent the features of a mention pair  $(m^a, m^b)$ . With this feature set, we found that the linear kernel is insufficient to fit the training data. However, using an rbf kernel would be too computationally expensive. Hence, we augment  $\hat{\phi}_{a,b}$  with a  $\hat{d}$ -dimensional feature vector  $[\psi^1 \cdots \psi^{\hat{d}}]$  to give a new feature vector

$$\phi_{a,b} = [\hat{\phi}_{a,b} \ \psi^1 \ \cdots \ \psi^{\hat{d}}], \tag{1}$$

where the  $\hat{d}$  augmented features  $[\psi^1 \cdots \psi^{\hat{d}}]$  are determined by

$$\psi^{j} = \exp(-\frac{\|\hat{\phi}_{a,b} - c^{j}\|^{2}}{d}), \forall j = 1, \dots, \hat{d}.$$
 (2)

Herein,  $c^1, \ldots, c^{\hat{d}}$  are the  $\hat{d}$  centroids of the randomly-selected subset C from all labeled mention pairs  $\{\hat{\phi}_{a,b} \mid (a,b) \in \mathcal{E}_1, \ldots, \mathcal{E}_N\}$ . In our experiments, we use the *k*-means algorithm to obtain the centroids of C.

**Ensemble** For domain v, we create a domainspecified ensemble  $\mathcal{F}^{(v)} = \{f^1, \ldots, f^\ell\}$  of  $\ell$  base models by including the closest-first and best-first mention-pair methods with the different C and t values. If multiple domains are provided, we gather all the domain-specific ensembles into a grand ensemble  $\mathcal{F} = \mathcal{F}^{(s_1)} \cup \cdots \mathcal{F}^{(s_p)} \cup \mathcal{F}^{(t)}$ .

### 3.2 Cross-domain Knowledge Learning

Generally, the feature distributions are different in different domains. Therefore, effective domain adaptation requires using some knowledge of crossdomain similarity. We now propose an approach to learn the parametric-distances between the documents in source and target domains to characterize this cross-domain knowledge.

**Distances between documents** A document  $D_i$  is represented by the sum of its new mention-pair features (Yu and Joachims, 2009; Finley and Joachims, 2005):

$$\Phi(\mathcal{D}_i) = \sum_{(a,b)\in\mathcal{E}_i} \phi_{a,b}.$$
(3)

The distance between a source labeled document  $\mathcal{D}_i^{(s_u)}$  in domain  $s_u$  and a target labeled document  $\mathcal{D}_i^{(t)}$  is parameterized as

$$\operatorname{Dist}(\mathcal{D}_i^{(s_u)}, \mathcal{D}_j^{(t)}; \boldsymbol{\mu}) = \boldsymbol{\mu}' \Delta(\mathcal{D}_i^{(s_u)}, \mathcal{D}_j^{(t)}), \quad (4)$$

where vector  $\boldsymbol{\mu} \in \mathbb{R}^{d+\hat{d}}$  is to be learned, and vector function  $\Delta(\mathcal{D}_i^{(s_u)}, \mathcal{D}_j^{(t)}) \in \mathbb{R}^{d+\hat{d}}$  is the Euclidean distance vector between two documents given by

$$\Delta(\mathcal{D}_i^{(s_u)}, \mathcal{D}_j^{(t)}) = (\Phi(\mathcal{D}_i^{(s_u)}) - \Phi(\mathcal{D}_j^{(t)}))$$
$$\odot (\Phi(\mathcal{D}_i^{(s_u)}) - \Phi(\mathcal{D}_j^{(t)})).$$
(5)

The operator  $\odot$  is the element-wise product. Distance (4) is actually the Mahanalobis distance (Yang and Jin, 2006) with the scaling of features:

$$(\Phi(\mathcal{D}_i^{(s_u)}) - \Phi(\mathcal{D}_j^{(t)}))'W(\Phi(\mathcal{D}_i^{(s_u)}) - \Phi(\mathcal{D}_j^{(t)})),$$

where W is a diagonal matrix with diagonal entries  $\mu$ . Matrix W is diagonal to reduce computation cost and to increase statistical confidence in estimation when there is limited target labeled data (as is typically the case in domain adaptation).

That  $\mu$  is the vector of diagonal entries in W requires that each entry in  $\mu$  is non-negative. If the  $l^{th}$  entry of  $\mu$  is non-zero, then the  $l^{th}$  feature in  $\phi_{a,b}$  contribute towards (4). To ensure that at least B features are used, we also constrain that each entry in  $\mu$  is not more than unity and that  $1'\mu \geq B$ .

**Matching best base models** For each labeled document  $\mathcal{D}_{j}^{(v)}$  in domain v, we identify the best performing base model  $f_{j}^{(v)^{*}}$  in  $\mathcal{F}^{(v)}$  with

$$f_j^{(v)^*} = \arg \max_{f \in \mathcal{F}^{(v)}} \Lambda(g(\mathcal{D}_j^{(v)}); f(\mathcal{D}_j^{(v)})), \quad (6)$$

where  $\Lambda(\cdot; \cdot)$  is the performance objective function to be instantiated in Section 3.3.

Then, for each source domain  $s_u$  and document  $\mathcal{D}_j^{(t)}$  in the target domain, we find the set  $\mathcal{I}(\mathcal{D}_j^{(t)}; s_u)$  of the documents in domain  $s_u$  that have the same best performing base model as that for  $\mathcal{D}_j^{(t)}$ :

$$\mathcal{I}(\mathcal{D}_{j}^{(t)}; s_{u}) = \{\mathcal{D}_{i}^{(s_{u})} \mid f_{i}^{(s_{u})^{*}} = f_{j}^{(t)^{*}}, \\ i = 1, \dots, N^{(s_{u})}\}.$$
(7)

The key idea in  $\mathcal{I}(\mathcal{D}_j^{(t)}; s_u)$  is to select documents in a source domain  $s_u$  that are similar to document  $D_j^{(t)}$  in the sense that they have the same best performing base model under a specific  $\Lambda$ . This ensures that optimization step (to be described next) is targeted towards  $\Lambda$  and not confounded by document pairs that should be disimilar anyway.

**Optimization** We determine the vector  $\mu$  by minimizing the parametric distance (4) between all target labeled documents and their corresponding source labeled document identified in the previous step. That is,

$$\min_{\boldsymbol{\mu}} \boldsymbol{\mu}' \sum_{j=1}^{N^{(t)}} \sum_{\mathcal{D}_i^{(s_u)} \in \mathcal{I}(\mathcal{D}_j^{(t)}; s_u)} \Delta(\mathcal{D}_i^{(s_u)}, \mathcal{D}_j^{(t)}).$$
(8)

The solution  $\mu$  to this linear programming problem can be regarded as the cross-domain knowledge between source domain  $s_u$  and the target domain t. Repeating for every source domain  $s_u$ ,  $u = 1, \ldots, p$ , gives the cross-domain knowledge between every source domain and the target domain.

The above three-steps procedure selects the effective features for each pair of source and target domains. Generally, the results of feature selection vary for different pairs of source and target domains, due to the diversities of the feature distributions in different domains.

#### **Decision Inference** 3.3

After ensemble creation and cross-domain knowledge learning, we need to provide the coreference result on an unseen document in the target domain based on the results of all the members in  $\mathcal{F}$ . Unlike the previous methods using the voting/average or their variants (Pang and Fan, 2009; Munson et al., 2005; Rahman and Ng, 2011a), we propose the following nearest neighbor based approach.

Given the grand ensemble  $\mathcal{F}$  and all labeled documents, the task is to predict on the target unlabeled document  $\mathcal{D}_{j}^{(t)}, j = 1, \dots, M$ . The idea of the proposed method is to first find the k most similar documents  $\mathcal{N}(\mathcal{D}_i^{(t)})$  from all labeled documents for document  $\mathcal{D}_{i}^{(t)}$ . Then, we choose the base model that performs best on the documents in  $\mathcal{N}(\mathcal{D}_i^{(t)})$  as the method  $f_j^{(t)^*}$  for document  $\mathcal{D}_j^{(t)}$ . Firstly, we employ the parametric-distance (4) to

measure the similarity between any labeled document  $\mathcal{D}_i^{(v)}, \forall v, i$ , from all source and target domains, and the target unlabeled document  $\mathcal{D}_{i}^{(t)}$ . Here, the cross-domain knowledge  $\mu$  in (4) has already been determined by the optimization (8) in Section 3.2.

Secondly, based on the computed distance values, we select k nearest neighbor documents for the target unlabeled document  $\mathcal{D}_i^{(t)}$  from all labeled documents  $\mathcal{D}_{i}^{(v)}, \forall v, i$ . These k nearest neighbor documents for document  $\mathcal{D}_{j}^{(t)}$  make up the set  $\mathcal{N}(\mathcal{D}_{j}^{(t)})$ . Thirdly, the optimal base model for the unlabeled

document  $\mathcal{D}_{i}^{(t)}$  prediction is chosen by

$$f_j^{(t)^*} = \arg \max_{\mathcal{D}_p \in \mathcal{N}(\mathcal{D}_j^{(t)}), f \in \mathcal{F}} \Lambda(g(\mathcal{D}_p); f(\mathcal{D}_p)).$$
(9)

We can instantiate the performance objective function  $\Lambda(q(\cdot); f(\cdot))$  in expressions (6) and (9) to be any coreference resolution measures, such as MUC, Rand index, B-CUBED and CEAF. We have not known of other (ensemble) coreference resolution methods that optimize for these measures. This absence is possibly due to their complex discrete and non-convex properties.

#### 3.4 Discussion

The above proposed adaptive ensemble approach incorporates the domain adaptation knowledge during

(a) the identification of similar documents between different domains and (b) the determination of active ensemble members. Beside these, it has the following features over other (ensemble) coreference methods: (i) It can optimize any user-specified objective measure via (6) and (9). An intuitive recommendation is to directly optimize for an objective function that matches the evaluation measure. (ii) It can make document-specific decisions, as expressions (4) and (9) deal with each testing document separately. (iii) The prediction on the testing document  $\mathcal{D}_{i}^{(t)}$  is not based on all members in  $\mathcal{F}$  but only on the active ensemble members  $\mathcal{N}(\mathcal{D}_j^{(t)})$ . This can filter out some potentially unsuitable base models for document  $\mathcal{D}_{i}^{(t)}$ . Moreover, the active ensemble members  $\mathcal{N}(\mathcal{D}_i^{(ec{t})})$  is dynamically adjusted for each test document.

For computational cost, the majority is by ensemble creation, since a large number of base models are usually used. This is common among all ensemble methods. In contrast, the costs in (4) and (9) are trivial as both are at the document level. The cost of generating centroids in (2) can also be high if the size of C is more than ten thousand, but this is still negligible compared to the cost of ensemble creation.

#### **Special Case: Within-domain Setting** 4

The adaptive ensemble method presented in Section 3 is for the domain adaptation setting. However, it is possible to simplify it for the special case of within-domain setting. In the within-domain setting, the adaptive ensemble method only has ensemble creation and decision inference steps.

In the ensemble creation step, we still use the closest-first and best-first mention-pair methods with various parameters to create the ensemble. Unlike the domain adaptation setting, here we can only use the labeled documents in the target domain to create the ensemble  $\mathcal{F}^{(t)}$ . Therefore, the size of ensemble here is reduced by p times compared to the domain adaptation setting.

In the decision inference step, we directly use the Euclidean distance  $\Delta(\mathcal{D}_i^{(t)},\mathcal{D}_j^{(t)})$  in (5) for the labeled document  $\mathcal{D}_i^{(t)}, i=1,\ldots,N^{(t)}$  and unlabeled document  $\mathcal{D}_{j}^{(t)}, j = 1, \dots, M$ . Based on these distance values, we similarly select k nearest neighbor documents  $\mathcal{N}(\mathcal{D}_{j}^{(t)})$  for document  $\mathcal{D}_{j}^{(t)}$ , and then determine the final method  $f_{j}^{(t)*}$  for document  $\mathcal{D}_{j}^{(t)}$  by (9) but with  $\mathcal{F}$  replaced by  $\mathcal{F}^{(t)}$ .

### **5** Experiments

We test the proposed adaptive method and several baselines under both the within-domain and the domain adaptation settings on the MUC-6 and ACE 2005 corpora. MUC-6 contains 60 documents. ACE 2005 contains 599 documents from six different domains: Newswire (NW), Broadcast News (BN), Broadcast Conversations (BC), Webblog (WL), Usenet (UN), and Conversational Telephone Speech (CTS). In all our experiments, we use two popular performance measures, B-CUBED Fmeasure (Bagga and Baldwin, 1998) and CEAF Fmeasure (Luo, 2005)<sup>1</sup>, to evaluate the coreference resolution result. Since the focus of the paper is to investigate the effectiveness of coreference resolution methods, we use the gold standard mentions in all experiments.

For the proposed method, the ensemble  $\mathcal{F}^{(v)}$  in every domain v has 208 members totally. They are created by the closest-first and the best-first mention-pair methods using SVM trained with parameter C taking values

$$C \in [0.001, 0.01, 0.1, 1, 10, 100, 1000, 1000]$$
 (10)

and using clustering with the threshold parameters t taking values

$$t \in [0.2, 0.25, 0.3, 0.34, 0.38, 0.4, 0.42, 0.44, 0.46, 0.48, 0.5, 0.6, 0.7].$$
(11)

The size of the selected subset C is fixed to 2000, and the number of centroids is determined by the validation procedure from four possible values [10, 20, 30, 40]. We use k-means algorithm to compute the centroids. Due to the randomness of subset C and k-means algorithm, we run the proposed method 5 times and report the average results. For the number of nearest neighbor k, we report three results, each for  $k \in \{1, 3, 5\}$ .

Table 1: The settings in the experiments under withindomain setting on MUC-6 and ACE 2005 corpora.  $N^{(t)}$ and  $M^{(t)}$  and Total are the numbers of training, testing and all documents respectively.

Domain	$N^{(t)}$	$M^{(t)}$	Total
MUC-6	30	30	60
BC	48	12	60
BN	181	45	226
CTS	31	8	39
NW	85	21	106
UN	39	10	49
WL	95	24	119

### 5.1 Within-domain Setting

We conduct the experiment under the within-domain setting on seven tasks, with the per-domain setting shown in Table 1. The validation set is created by further splitting training data into validation training and validation testing sets with the ratio of  $\frac{N^{(t)}}{M^{(t)}}$ , where  $N^{(t)}$  and  $M^{(t)}$  are given in Table 1. In this experiment, we attempt to study the following three things. First, we investigate whether the proposed ensemble method is better than the tuned mentionpair methods and other ensemble methods. Second, we investigate the optimal number of active ensemble members. Third, we investigate the impact to the performance of the coreference system, when different evaluation measures.

For the proposed ensemble method, we experimented with nearest neighbor set of sizes k = 1, 3, 5paired with objective function  $\Lambda$  in (9) set to Rand Index, CEAF or B-CUBED. For baselines, the following four are used:

- Two mention-pair baselines. Two baselines are the closest-first and the best-first mention-pair methods with the tuned parameters C and t. In the tuning process, the ranges of C and t are specified in (10) and (11) respectively. These two mention-pair methods are named as  $S_c$  and  $S_b$  for short.
- Two existing ensemble baselines. The other two baselines are the ensemble methods using the voting procedure in decision inference.

<sup>&</sup>lt;sup>1</sup>More exactly, we use the widely used  $\phi_3$ -CEAF F-measure.

	Baselines			$\Lambda = Rand$			$\Lambda = \mathrm{CEAF}$			$\Lambda = \text{B-CUBED}$			
	$\mathbf{S}_{c}$	$\mathbf{S}_b$	$E_m$	$E_c$	k=1	3	5	k=1	3	5	k=1	3	5
MUC-6	66.1	66.1	61.9	57.1	67.6	67.3	68.5	65.2	64.1	65.5	68.7	66.7	67.5
BC	64.1	65.1	34.2	24.8	65.5	65.4	65.7	65.9	65.5	62.9	66.5	66.1	66.0
BN	75.9	74.8	57.7	48.0	75.7	75.1	74.9	76.3	75.9	75.3	76.4	76.3	76.7
CTS	71.0	65.1	39.6	31.5	70.6	69.3	68.3	71.3	69.9	70.4	71.7	70.6	69.1
NW	74.6	74.4	45.6	34.1	74.3	74.8	72.9	73.2	71.4	70.1	75.0	74.6	73.7
UN	69.5	70.2	44.1	27.4	70.4	69.9	69.3	69.6	67.6	66.0	70.3	71.4	70.3
WL	73.8	75.4	69.8	58.5	75.5	74.6	73.9	75.5	73.0	73.4	76.2	75.5	75.6
Average	70.7	70.2	50.4	40.2	71.4	70.9	70.5	71.0	69.6	69.1	72.1	71.6	71.3

Table 2: B-CUBED F-measure results by all methods under within-domain setting on MUC-6 and ACE 2005 corpora.

These two baselines use the same ensemble as the proposed method for fair comparison. In decision inference, these two baselines use the mention-based voting and cluster-based voting respectively, as proposed in (Rahman and Ng, 2011a). In these two baselines, all members in the ensemble participate the voting process. These two ensemble baselines are named as  $E_m$ and  $E_c$  for short.

Tables 2 and 3 show the experiment results using B-CUBED and CEAF as the evaluation measures respectively. The best result for each of the seven tasks is highlighted in bold. The last rows of the tables show the average performance value among all seven tasks.

From the results, we observe that the proposed ensemble method with objective function matching the evaluation measure and with k = 1 generally performs best among all methods and all tasks. Surprisingly, the common ensemble method with mentionbased voting  $E_m$  and cluster-based voting  $E_c$  strategies do not perform well. The plausible reason is the current ensemble may incorporate some bad base models due to inappropriate C and t values, which would undermine the voting result. Nevertheless, it is difficult to judge the quality of the ensemble members in advance. Therefore, this validates the importance of choosing an active set of ensemble members in decision inference. The better performance of the proposed method over the mention-pair baselines  $S_c$ and  $S_b$  is probably because of the document-specific decision. This is reasonable, as different base models in the ensemble would be good at predicting the different documents. For the proposed ensemble method with various configurations, we observe using an objective function that matches the evaluation measures is generally better. An exception is the MUC-6 and BN tasks in CEAF F-measure. We also observe that the ensemble method with k = 1is generally better than that with the larger k, except the BN and UN tasks in B-CUBED F-measure. This suggests that the fewer the active ensemble members the better the generalization performance. Following (Rahman and Ng, 2011a), we also conduct the Student's t-test, and the results show that the proposed method with the objective function matching the evaluation measure and with k = 1 is significantly better than the best baseline. In contrast, the two baseline ensemble methods that use voting are significantly worse than the best baseline. The significance level 0.05.

### 5.2 Domain-adaptation Setting

We employ ACE 2005 corpora to simulate the domain adaptation settings in experiments. Specifically, we create six domain adaptation tasks, BC, BN, CTS, NW, UN, WL in total. Each task has one target domain and five source domains. For example, in the task UN, the target domain is UN while the other five source domains are BC, BN, CTS, NW and WL. The number of labeled documents in each domain is as the same as in Table 1, except when that domain is the target domain, in which case we use only five labeled documents. The number of test

	Baselines			$\Lambda = Rand$			$\Lambda = \text{B-CUBED}$			$\Lambda = \mathrm{CEAF}$			
	$\mathbf{S}_{c}$	$\mathbf{S}_b$	$E_m$	$E_c$	k=1	3	5	k=1	3	5	k=1	3	5
MUC-6	62.6	62.5	62.7	57.5	62.0	60.6	61.0	64.5	62.7	63.8	63.1	58.7	59.2
BC	58.8	56.5	36.6	26.6	56.7	57.1	57.0	58.3	58.8	57.2	59.3	59.2	58.4
BN	67.9	66.5	55.1	44.7	69.4	69.4	69.9	69.8	70.2	69.6	69.5	69.0	68.7
CTS	61.0	60.7	38.6	31.5	67.1	66.9	63.6	68.1	68.4	68.2	68.5	67.6	67.7
NW	66.9	66.4	41.1	31.2	68.4	68.0	64.6	69.2	68.4	66.4	69.3	66.1	66.7
UN	62.5	63.5	46.2	28.9	62.9	61.8	60.9	62.2	63.7	62.9	63.9	61.5	60.4
WL	69.7	70.3	63.5	54.3	70.7	70.2	72.5	71.5	71.4	72.3	72.4	69.4	70.0
Average	64.2	63.8	49.1	39.2	65.3	64.9	64.2	66.2	66.2	65.8	66.6	64.5	64.5

Table 3: CEAF F-measure results by all methods under within-domain setting on MUC-6 and ACE 2005 corpora.

(or unlabeled) documents in the target document is also the same as in Table 1. The validation set is created similarly as in the experiment under withindomain setting.

For the proposed ensemble method, we heuristically determine the parameter B in  $\mu$  to be the number of non-zero elements in  $\Gamma$ , where

$$\Gamma = \sum_{j=1}^{N^{(t)}} \sum_{\mathcal{D}_i^{(s_u)} \in \mathcal{I}(\mathcal{D}_j^{(t)}; s_u)} \Delta(\mathcal{D}_i^{(s_u)}, \mathcal{D}_j^{(t)}).$$

Making use of the conclusion in the experiments for the within-domain setting, we fix the optimized measure to be the final performance measure in (9). We compare with the following five baselines.

- Two mention-pair baselines in within-domain setting. Two baselines are same as S<sub>c</sub> and S<sub>b</sub> in the experiments under within-domain settings, except that the labeled training documents are reduced to 5.
- Three proposed adaptive ensemble methods without cross-domain knowledge learning. These three baselines uses neighborhood sizes k = 1, 3, 5 with the grand ensemble  $\mathcal{F}$  rather than the target domain ensemble  $\mathcal{F}^{(t)}$ . In another words, these three baselines are the same as the proposed method, but with  $\mu = 1$ .

Tables 4 and 5 show the experimental results in the domain adaptation settings using B-CUBED and

CEAF as the final performance measures respectively. From the results, we can see that the proposed method with cross-domain knowledge generally outperforms all the five baselines. Among them, the best proposed domain adaptation method on average outperforms the best of  $S_c$ ,  $S_b$  by 7.2% for B-CUBED F-measure and 3% for CEAF F-measure. The grand-ensemble baselines are also significantly better than the within-domain baselines. These results clearly illustrate the usefulness of making use of the labeled documents in the source domains. For the comparison between the proposed method with and without cross-domain knowledge learning, all tasks, except UN task in CEAF F-measure, show the superiority of the proposed method with crossdomain knowledge learning. Among them, except tasks BN and CTS in B-CUBED F-measure, the performance gains are among 1%-3% for all tasks in both measures. These results verify the necessity of cross-domain knowledge learning. For the comparison of the proposed method with different k, unlike the results in the within-domain setting, the results here show that choosing optimal k is taskdependent. The reason of this observation is not clear yet. It is plausible due to the increased uncertainties from multiple domains.

# 6 Conclusions and Future Work

In this paper, we proposed an adaptive ensemble method for coreference resolution under both within-domain and domain adaptation settings. The key advantage of the proposed method is incor-

	Within-domain		Grar	nd ense	mble	Domain-adaptation		
	$\mathbf{S}_{c}$	$\mathbf{S}_b$	k=1	3	5	k=1	3	5
BC	58.0	65.1	65.0	67.1	67.0	67.5	68.2	67.7
BN	72.7	73.8	75.0	75.3	75.0	75.3	75.4	74.3
CTS	63.2	62.1	65.7	64.8	64.0	64.1	65.8	65.8
NW	54.9	54.6	73.6	73.1	74.2	73.0	74.4	74.7
UN	66.5	42.7	67.2	68.2	68.9	<b>69.7</b>	68.7	68.2
WL	68.6	73.2	73.0	72.6	73.4	74.8	74.5	73.6
Average	64.0	61.9	69.9	70.2	70.4	70.7	71.2	70.7

Table 4: B-CUBED F-measure results by all methods under domain adaptation setting on ACE 2005 corpora, with  $\Lambda$  set to B-CUBED. The *within-domain* and *grand ensemble* methods are the baselines.

Table 5: CEAF F-measure results by all methods under domain adaptation setting on ACE 2005 corpora, with  $\Lambda$  set to CEAF. The *within-domain* and *grand ensemble* methods are the baselines.

	Within	n-domain	Gran	nd ensei	mble	Domain-adaptation		
	$\mathbf{S}_{c}$	$\mathbf{S}_b$	k=1	3	5	k=1	3	5
BC	55.7	43.7	56.9	57.6	57.3	58.5	58.8	57.2
BN	65.8	67.2	65.9	64.1	65.8	63.9	62.7	67.2
CTS	56.0	51.0	56.6	54.6	53.7	58.6	57.4	55.3
NW	52.7	55.0	66.4	64.1	63.8	69.4	66.7	66.8
UN	64.0	39.1	63.6	63.7	64.4	64.3	62.9	62.7
WL	70.3	64.2	68.1	67.8	70.2	67.3	69.6	72.0
Average	60.7	53.4	62.9	62.0	62.5	63.7	63.0	63.5

porating the cross-domain knowledge to aid coreference resolution learning. This is useful when the labeled coreference labels are scarce. We also demonstrate that the proposed adaptive ensemble method can be readily applied to conventional coreference tasks without cross-domain knowledge learning. Compared with existing ensemble methods, the proposed method is simultaneously endowed with the following three distinctive features: optimizing any user-specified performance measure, making the document-specific prediction and automatically adjusting the active ensemble members. In the experiments under both within-domain settings and domain adaptation settings, the results evidence the effectiveness of the proposed cross-domain knowledge learning method, and also demonstrate the superiority of the proposed adaptive ensemble method over other baselines.

Currently, the proposed method relies on some

limited target annotations. It would be interesting to consider the pure unsupervised tasks that have no any target annotations. Besides, to develop some better ways for document-level representation, e.g., incorporating the domain knowledge, also deserves our attentions. Similarly, to extend the diagonal Mahalanobis matrix to the general covariance matrix is also desirable. Last but not least, to find a more systematical way to determine the optimal k in the proposed method is also our possible future work.

### Acknowledgments

This work is supported by DSO grant DSOCL10021.

### References

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space

model. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1, ACL'98, pages 79–85.

- Christopher M. Bishop. 2007. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, 1st ed. 2006. corr. 2nd printing edition, October.
- John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of EMNLP*, pages 120–128.
- Leo Breiman. 1996. Bagging predictors. *Machine Learning*, 24(2):123–140, August.
- Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32, October.
- Pascal Denis and Jason Baldridge. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *Proc HLT*, pages 236– 243, Rochester, New York, April.
- Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.
- Thomas Finley and Thorsten Joachims. 2005. Supervised clustering with support vector machines. In *Proc. ICML*.
- Yoav Freund and Robert E. Schapire. 1996. Experiments with a New Boosting Algorithm. In *Proc. ICML*, pages 148–156.
- Honglei Guo, Huijia Zhu, Zhili Guo, Xiaoxun Zhang, Xian Wu, and Zhong Su. 2009. Domain adaptation with latent semantic association for named entity recognition. NAACL '09, pages 281–289.
- Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proc. ACL*, pages 264–271, Prague, Czech Republic, June.
- Jing Jiang. 2009. Multi-task transfer learning for weakly-supervised relation extraction. 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, pages 1012–1020, Suntec, Singapore, August. Association for Computational Linguistics.
- Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, pages 25– 32.
- Art Munson, Claire Cardie, and Rich Caruana. 2005. Optimizing to arbitrary NLP metrics using ensemble selection. In *Proc HLT and EMNLP*, pages 539–546.

- Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proc. ACL*, pages 104–111.
- Vincent Ng and Claire Cardie. 2003. Weakly supervised natural language learning without redundant views. In *Proc. HLT-NAACL*.
- Vincent Ng. 2005. Machine learning for coreference resolution: From local classification to global ranking. In *Proceedings of the ACL*, pages 157–164.
- Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October.
- Wenbo Pang and Xiaozhong Fan. 2009. Chinese coreference resolution with ensemble learning. In *Proc. PACIIA*, pages 236–243.
- Altaf Rahman and Vincent Ng. 2011a. Ensemblebased coreference resolution. In *Proceedings of IJ-CAI*, pages 1884–1889.
- Altaf Rahman and Vincent Ng. 2011b. Narrowing the modeling gap: A cluster-ranking approach to coreference resolution. *JAIR*, 1:469–52.
- William M. Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):pp. 846–850.
- W. M. Soon, H. T. Ng, and D. C. Y. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, pages 521– 544.
- Veselin Stoyanov, Claire Cardie, Nathan Gilbert, Ellen Riloff, David Buttler, and David Hysom. 2010. Coreference resolution with reconcile. In *Proc. ACL*, pages 156–161.
- Songbo Tan, Yuefen Wang, Gaowei Wu, and Xueqi Cheng. 2008. Using unlabeled data to handle domaintransfer problem of semantic detection. In Proceedings of the 2008 ACM symposium on Applied computing, SAC '08, pages 896–903.
- S. Vemulapalli, X. Luo, J.F.Pitrelli, and I. Zitouni. 2009. classifier combination applied to coreference resolution. In *NAACL HLT Student Rsearch Workshop*.
- Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A modeltheoretic coreference scoring scheme. In Proceedings of the 6th conference on Message understanding, MUC6 '95, pages 45–52.
- Liu Yang and Rong Jin. 2006. Distance Metric Learning: A Comprehensive Survey. Technical report, Department of Computer Science and Engineering, Michigan State University.
- Chun-Nam John Yu and Thorsten Joachims. 2009. Learning structural SVMs with latent variables. In *Proc. ICML*, pages 1169–1176, New York, NY, USA.