# Domain Adaptation with Active Learning for Coreference Resolution

**Shanheng Zhao**
Elance
441 Logue Ave
Mountain View, CA 94043, USA
szhao@elance.com

**Hwee Tou Ng**
Department of Computer Science
National University of Singapore
13 Computing Drive, Singapore 117417
nght@comp.nus.edu.sg

## Abstract

In the literature, most prior work on coreference resolution centered on the newswire domain. Although a coreference resolution system trained on the newswire domain performs well on newswire texts, there is a huge performance drop when it is applied to the biomedical domain. In this paper, we present an approach integrating domain adaptation with active learning to adapt coreference resolution from the newswire domain to the biomedical domain. We explore the effect of domain adaptation, active learning, and target domain instance weighting for coreference resolution. Experimental results show that domain adaptation with active learning and target domain instance weighting achieves performance on MEDLINE abstracts similar to a system trained on coreference annotation of only target domain training instances, but with a greatly reduced number of target domain training instances that we need to annotate.

## 1 Introduction

Coreference resolution is the task of determining whether two or more noun phrases (NPs) in a text refer to the same entity. Successful coreference resolution benefits many natural language processing (NLP) tasks, such as information extraction and question answering. In the literature, most prior work on coreference resolution recasts the problem as a two-class classification problem. Machine learning-based classifiers are applied to determine whether a candidate anaphor and a potential antecedent are coreferential (Soon et al., 2001; Ng and Cardie, 2002; Stoyanov et al., 2009; Zhao and Ng, 2010).

In recent years, with the advances in biological and life science research, there is a rapidly increasing number of biomedical texts, including research papers, patent documents, etc. This results in an increasing demand for applying natural language processing and information retrieval techniques to efficiently exploit information contained in these large amounts of texts. However, coreference resolution, one of the core tasks in NLP, has only a relatively small body of prior research in the biomedical domain (Kim et al., 2011a; Kim et al., 2011b).

A large body of prior research on coreference resolution focuses on texts in the newswire domain. Standardized data sets, such as MUC (DARPA Message Understanding Conference, (MUC-6, 1995; MUC-7, 1998)) and ACE (NIST Automatic Content Extraction Entity Detection and Tracking task, (NIST, 2002)) data sets are widely used in the study of coreference resolution.

Traditionally, in order to apply supervised machine learning approaches to an NLP task in a specific domain, one needs to collect a text corpus in the domain and annotate it to serve as training data. Compared to other NLP tasks, e.g., part-of-speech (POS) tagging or named entity (NE) tagging, the annotation for coreference resolution is much more challenging and time-consuming. The reason is that in tasks like POS tagging, an annotator only needs to focus on each markable (a word, in the case of POS tagging) and a small window of its neighboring words. In contrast, to annotate a coreferential relation, an annotator needs to first recognize whether a certain text span is a markable, and then scan through the text preceding the markable (a potential anaphor) to look for the antecedent. It also requires the annotator to understand the text in order to annotate coreferential relations, which are *semantic* in nature. If a markable is non-anaphoric, the annotator has to scan to the beginning of the text to realize that. Cohen et al. (2010) reported that it took an average of 20 hours to annotate coreferential relations in a single

*Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) @ EACL 2014*, pages 21–29,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

document with an average length of 6,155 words, while an annotator could annotate 3,000 words per hour in POS tag annotation (Marcus et al., 1993).

The simplest approach to avoid the time-consuming data annotation in a new domain is to train a coreference resolution system on a resource-rich domain and apply it to a different target domain without any additional data annotation. Although coreference resolution systems work well on test texts in the same domain as the training texts, there is a huge performance drop when they are tested on a different domain. This motivates the usage of domain adaptation techniques for coreference resolution: adapting a coreference resolution system from one source domain in which we have a large collection of annotated data, to a second target domain in which we need good performance. It is almost inevitable that we annotate *some* data in the target domain to achieve good coreference resolution performance. The question is how to minimize the amount of annotation needed. In the literature, active learning has been exploited to reduce the amount of annotation needed (Lewis and Gale, 1994). In contrast to annotating the entire data set, active learning selects only a subset of the data to annotate in an iterative process. How to apply active learning and integrate it with domain adaptation remains an open problem for coreference resolution.

In this paper, we explore domain adaptation for coreference resolution from the resource-rich newswire domain to the biomedical domain. Our approach comprises domain adaptation, active learning, and target domain instance weighting to leverage the existing annotated corpora from the newswire domain, so as to reduce the cost of developing a coreference resolution system in the biomedical domain. Our approach achieves comparable coreference resolution performance on MEDLINE abstracts, but with a large reduction in the number of training instances that we need to annotate. To the best of our knowledge, our work is the first to combine domain adaptation and active learning for coreference resolution.

The rest of this paper is organized as follows. We first review the related work in Section 2. Then we describe the coreference resolution system in Section 3, and the domain adaptation and active learning techniques in Section 4. Experimental results are presented in Section 5. Finally, we analyze the results in Section 6 and conclude in Section 7.

## 2 Related Work

Not only is there a relatively small body of prior research on coreference resolution in the biomedical domain, there are also fewer annotated corpora in this domain. Castaño et al. (2002) were among the first to annotate coreferential relations in the biomedical domain. Their annotation only concerned the pronominal and nominal anaphoric expressions in 46 biomedical abstracts. Gasperin and Briscoe (2007) annotated coreferential relations on 5 full articles in the biomedical domain, but only on noun phrases referring to bio-entities. Yang et al. (2004) annotated full NP coreferential relations on biomedical abstracts of the GENIA corpus. The ongoing project of the CRAFT corpus is expected to annotate all coreferential relations on full text of biomedical articles (Cohen et al., 2010).

Unlike the work of (Castaño et al., 2002), (Gasperin and Briscoe, 2008), and (Gasperin, 2009) that resolved coreferential relations on certain restricted entities in the biomedical domain, we resolve all NP coreferential relations. Although the GENIA corpus contains 1,999 biomedical abstracts, Yang et al. (2004) tested only on 200 abstracts under 5-fold cross validation. In contrast, we randomly selected 399 abstracts in the 1,999 MEDLINE abstracts of the GENIA-MEDCo corpus as the test set, and as such our evaluation was carried out on a larger scale.

Domain adaptation has been studied and successfully applied to many natural language processing tasks (Jiang and Zhai, 2007; Daume III, 2007; Dahlmeier and Ng, 2010; Yang et al., 2012). On the other hand, active learning has also been applied to NLP tasks to reduce the need of data annotation in the literature (Tang et al., 2002; Laws et al., 2012; Miller et al., 2012). Unlike the aforementioned work that applied only one of domain adaptation or active learning to NLP tasks, we combine both. There is relatively less research on combining domain adaptation and active learning together for NLP tasks (Chan and Ng, 2007; Zhong et al., 2008; Rai et al., 2010). Chan and Ng (2007) and Zhong et al. (2008) used *count merging* and *augment*, respectively, as their domain adaptation techniques whereas we apply and compare multiple state-of-the-art domain adaptation techniques. Rai et al. (2010) exploited a streaming active learning setting whereas ours is pool-based.

Dahlmeier and Ng (2010) evaluated the performance of three previously proposed domain adaptation algorithms for semantic role labeling. They evaluated the performance of domain adaptation with different sizes of target domain training data. In each of their experiments with a certain target domain training data size, the target domain training data were added all at once. In contrast, we add the target domain training instances selectively in an iterative process. Different from (Dahlmeier and Ng, 2010), we weight the target domain instances to further boost the performance of domain adaptation. Our work is the first systematic study of domain adaptation with active learning for coreference resolution. Although Gasperin (2009) tried to apply active learning for anaphora resolution, her results were negative: using active learning was not better than randomly selecting instances in her work. Miwa et al. (2012) incorporated a rule-based coreference resolution system for automatic biomedical event extraction, and showed that by adding training data from other domains as supplementary training data and using domain adaptation, one can achieve a higher F-measure in event extraction.

## 3 Coreference Resolution

The gold standard annotation and the output by a coreference resolution system are called the key and the response, respectively. In both the key and the response, a coreference chain is formed by a set of coreferential markables. A *markable* is a noun phrase which satisfies the markable definition in an individual corpus. Here is an example:

> When ***the same MTHC lines*** are exposed to TNF-alpha in combination with IFN-gamma, ***the cells*** instead become DC.

In the above sentence, ***the same MTHC lines*** and ***the cells*** are referring to the same entity and hence are coreferential. It is possible that more than two markables are coreferential in a text. The task of coreference resolution is to determine these relations in a given text.

To evaluate the performance of coreference resolution, we follow the MUC evaluation metric introduced by (Vilain et al., 1995). Let $S_i$ be an equivalence class generated by the key (i.e., $S_i$ is a coreference chain), and $p(S_i)$ be a partition of $S_i$ relative to the response. Recall is the number of correctly identified links over the number of links in the key: $Recall = \frac{\sum(|S_i|-|p(S_i)|)}{\sum(|S_i|-1)}$. Precision, on the other hand, is defined in the opposite way by switching the role of key and response. F-measure is a trade-off between recall and precision: $F = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$.

## 4 Domain Adaptation with Active Learning

### 4.1 Domain Adaptation

Domain adaptation is applicable when one has a large amount of annotated training data in the source domain and a small amount or none of the annotated training data in the target domain. We evaluate the AUGMENT technique introduced by (Daume III, 2007), as well as the INSTANCE WEIGHTING (IW) and the INSTANCE PRUNING (IP) techniques introduced by (Jiang and Zhai, 2007).

#### 4.1.1 AUGMENT

Daume III (2007) introduced a simple domain adaptation technique by feature space augmentation. It maps the feature space of each instance into a feature space of higher dimension. Suppose $x$ is the feature vector of an instance. Define $\Phi^s$ and $\Phi^t$ to be the mappings of an instance from the original feature space to an augmented feature space in the source and the target domain, respectively:

$$\Phi^s(x) = \langle x, x, \mathbf{0} \rangle \quad (1)$$

$$\Phi^t(x) = \langle x, \mathbf{0}, x \rangle \quad (2)$$

where $\mathbf{0} = \langle 0, 0, \ldots, 0 \rangle$ is a zero vector of length $|x|$. The mapping can be treated as taking each feature in the original feature space and making three versions of it: a general version, a source-specific version, and a target-specific version. The augmented source domain data will contain only the general and the source-specific versions, while the augmented target domain data will contain only the general and the target-specific versions.

#### 4.1.2 INSTANCE WEIGHTING and INSTANCE PRUNING

Let $x$ and $y$ be the feature vector and the corresponding true label of an instance, respectively.

Jiang and Zhai (2007) pointed out that when applying a classifier trained on a source domain to a target domain, the joint probability $P_t(x, y)$ in the target domain may be different from the joint probability $P_s(x, y)$ in the source domain. They proposed a general framework to use $P_s(x, y)$ to estimate $P_t(x, y)$. The joint probability $P(x, y)$ can be factored into $P(x, y) = P(y|x)P(x)$. The adaptation of the first component is labeling adaptation, while the adaptation of the second component is instance adaptation. We explore only labeling adaptation.

To calibrate the conditional probability $P(y|x)$ from the source domain to the target domain, ideally each source domain training instance $(x_i, y_i)$ should be given a weight $\frac{P_t(y_i^s|x_i^s)}{P_s(y_i^s|x_i^s)}$. Although $P_s(y_i^s|x_i^s)$ can be estimated from the source domain training data, the estimation of $P_t(y_i^s|x_i^s)$ is much harder. Jiang and Zhai(2007) proposed two methods to estimate $P_t(y_i^s|x_i^s)$: Instance Weighting and Instance Pruning. Both methods first train a classifier with a small amount of target domain training data. Then, Instance Weighting directly estimates $P_t(y_i^s|x_i^s)$ using the trained classifier. Instance Pruning, on the other hand, removes the top $N$ source domain instances that are predicted wrongly, ranked by the prediction confidence.

#### 4.1.3 Target Domain Instance Weighting

Both Instance Weighting and Instance Pruning set the weights of the source domain instances. In domain adaptation, there are typically many more source domain training instances than target domain training instances. Target domain instance weighting can effectively reduce the imbalance. Unlike Instance Weighting and Instance Pruning in which each source domain instance is weighted individually, we give all target domain instances the same weight. This target domain instance weighting scheme is not only complementary to Instance Weighting and Instance Pruning, but is also applicable to Augment.

### 4.2 Active Learning

Active learning iteratively selects the most informative instances to label, adds them to the training data pool, and trains a new classifier with the enlarged data pool. We follow (Lewis and Gale, 1994) and use the uncertainty sampling strategy in our active learning setting.

```
$D_s \leftarrow$ the set of source domain training instances
$D_t \leftarrow$ the set of target domain training instances
$D_a \leftarrow \emptyset$
$\Gamma \leftarrow$ coreference resolution system trained on $D_s$
$T \leftarrow$ number of iterations
for i from 1 to T do
    for each $d_i \in D_t$ do
        $\widehat{d_i} \leftarrow$ prediction of $d_i$ using $\Gamma$
        $p_i \leftarrow$ prediction confidence of $\widehat{d_i}$
    end for
    $D'_a \leftarrow$ top $N$ instances with the lowest $p_i$
    $D_a \leftarrow D_a + D'_a$
    $D_t \leftarrow D_t - D'_a$
    provide correct labels to the unlabeled instances in $D'_a$
    $\Gamma \leftarrow$ coreference resolution system trained on $D_s$ and
    $D_a$ using the chosen domain adaptation technique
end for
```

Figure 1: An algorithm for domain adaptation with active learning

### 4.3 Domain Adaptation with Active Learning

Combining domain adaptation and active learning together, the algorithm we use is shown in Figure 1.

In our domain adaptation setting, there is a parameter $\lambda_t$ for target domain instance weighting. Because the number of target domain instances is different in each iteration, the weight should be adjusted in each iteration. We give all target domain training instances an equal weight of $\lambda_t = N_s/N_t$, where $N_s$ and $N_t$ are the numbers of instances in the source domain and the target domain in the current iteration, respectively. We set $N = 10$ to add 10 instances in each iteration to speed up the active learning process.

To provide the correct labels, the labeling process shows the text on the screen, highlights the two NPs, and asks the annotator to decide if they are coreferential. In our experiments, this is simulated by providing the gold standard coreferential information on this NP pair to the active learning process.

## 5 Experiments

### 5.1 The Corpora

We explore domain adaptation from the newswire domain to the biomedical domain. The newswire and biomedical domain data that we use are the ACE Phase-2 corpora and the GENIA-MEDCo corpus, respectively. The ACE corpora contain 422 and 92 training and test texts, respectively (NIST, 2002). The texts come from three newswire sources: BNEWS, NPAPER, and NWIRE. The GENIA-MEDCo corpus contains 1,999 MEDLINE abstracts[1]. We randomly split the GENIA corpus into a training set and a test set, containing 1,600 and 399 texts, respectively.

### 5.2 The Coreference Resolution System

In this study, we use Reconcile, a state-of-the-art coreference resolution system implemented by (Stoyanov et al., 2009). The input to the coreference resolution system is raw text, and we apply a sequence of preprocessing components to process it. Following Reconcile, the individual preprocessing steps include: 1) sentence segmentation (using the OpenNLP toolkit[2]); 2) tokenization (using the OpenNLP toolkit); 3) POS tagging (using the OpenNLP toolkit); 4) syntactic parsing (using the Berkeley Parser[3]); and 5) named entity recognition (using the Stanford NER[4]). Markables are extracted as defined in each individual corpus. All possible markable pairs in the training and test set are extracted to form training and test instances, respectively. The learning algorithm we use is maximum entropy modeling, implemented in the DALR package[5] (Jiang and Zhai, 2007). The coreference resolution system employs a comprehensive set of 62 features to represent each training and test instance, including lexical, proximity, grammatical, and semantic features (Stoyanov et al., 2009). We do not introduce additional features motivated from the biomedical domain, but use the same feature set for both the source and target domains.

### 5.3 Preprocessing

For the ACE corpora, all preprocessing components use the original models (provided by the OpenNLP toolkit, the Berkeley Parser, and the Stanford NER). For the GENIA corpus, since it is from a very different domain, the original models do not perform well. However, the GENIA corpus contains multiple layers of annotations. We use these annotations to re-train each of the preprocessing components (except tokenization) using the 1,600 training texts of the GENIA corpus[6]. We do not use any texts from the test set when training these models. Also, we do not use any NLP toolkits from the biomedical domain, but only use general toolkits trained with biomedical training data. These re-trained preprocessing components are then applied to process the entire GENIA corpus, including both the training and test sets.

| | NPAPER TRAIN | NPAPER TEST | GENIA TRAIN | GENIA TEST |
|---|---|---|---|---|
| Number of Docs | | | | |
| | 76 | 17 | 1,600 | 399 |
| Number of Words | | | | |
| Total | 68,463 | 17,350 | 391,380 | 95,405 |
| Avg. | 900.8 | 1,020.6 | 244.6 | 239.1 |
| Number of Markables | | | | |
| Total | 21,492 | 5,153 | 99,408 | 24,397 |
| Avg. | 282.8 | 303.1 | 62.1 | 61.1 |
| Number of Instances | | | | |
| Total | 3,365,680 | 871,314 | 3,335,640 | 798,844 |
| Avg. | 44,285.3 | 51,253.8 | 2,084.8 | 2,002.1 |

Table 1: Statistics of the NPAPER and GENIA data sets

Instead of using the entire ACE corpora, we choose the NPAPER portion of the ACE corpora as the source domain in the experiments, because it is the best performing one among the three portions. Under these preprocessing settings, the recall percentages of markable extraction on the training and test set of the NPAPER corpus are 94.5% and 95.5% respectively, while the recall percentages of markable extraction on the training and test set of the GENIA corpus are 87.6% and 86.6% respectively. The statistics of the NPAPER and the GENIA corpora are listed in Table 1.

### 5.4 Baseline Results

Under our experimental settings, a coreference resolution system that is trained on the NPAPER training set and tested on the NPAPER test set achieves recall, precision, and F-measure of 59.0%, 70.6%, and 64.3%, respectively. This is comparable to the state-of-the-art performance (Stoyanov et al., 2009). Table 2 compares the performance of testing on the GENIA test set, but training with the GENIA training set or the NPAPER training set. Training with in-domain data achieves an F-measure that is 9.1% higher than training with out-of-domain data. Training with

[1] http://nlp.i2r.a-star.edu.sg/medco.html

[2] http://opennlp.sourceforge.net/

[3] http://code.google.com/p/berkeleyparser/

[4] http://nlp.stanford.edu/ner/

[5] http://www.mysmu.edu/faculty/jingjiang/software/DALR.html

[6] It turned out that the re-trained tokenization model gave poorer performance and produced many errors on punctuation symbols. Thus, we stuck to using the original tokenization model.

| Training Set | Recall | Precision | F-measure |
|---|---|---|---|
| GENIA Training Set | 37.7 | 71.9 | 49.5 |
| NPAPER Training Set | 30.3 | 60.7 | 40.4 |

Table 2: MUC F-measures on the GENIA test set

in-domain data is better than training with out-of-domain data for both recall and precision. This confirms the impact of domain difference between the newswire and the biomedical domain.

### 5.5 Domain Adaptation with Active Learning

In the experiments on domain adaptation with active learning for coreference resolution, we assume that the source domain training data are annotated. The target domain training data are *not* annotated but are used as a data pool for instance selection. The algorithm selects the instances in the data pool to annotate and add them to the training data to update the classifier. The target domain test set is strictly separated from this data pool, i.e., none of the target domain test data are used in the instance selection process of active learning.

From Table 1, one can see that both training sets in the NPAPER and the GENIA corpora contain large numbers of training instances. Instead of using the entire training sets in the experiments, we use a smaller subset due to several reasons. First, to train a coreference resolution classifier, we do not need so much training data (Soon et al., 2001). Second, a large number of training instances will slow the active learning process. Third, a smaller source domain training corpus suggests a more modest annotation effort even on the source domain. Lastly, a smaller target domain training corpus means that fewer words need to be read by human annotators to label the data.

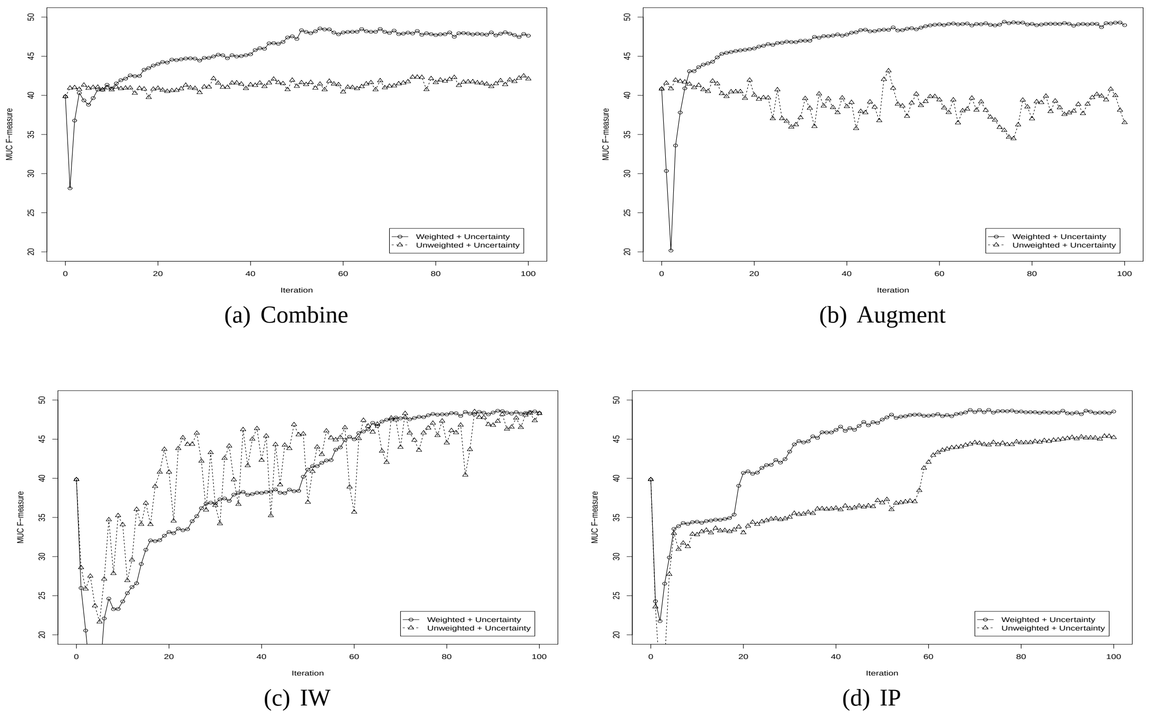
We randomly choose 10 NPAPER texts as the source domain training set. A coreference resolution system that is trained on these 10 texts and tested on the entire NPAPER test set achieves recall, precision, and F-measure of 60.3%, 70.6%, and 65.0%, respectively. This is comparable to (actually slightly better than) a system trained on the entire NPAPER training set. As for the GENIA training set, we randomly choose 40 texts as the target domain training data. To avoid selection bias, we perform 5 random trials, i.e., choosing 5 sets, each containing 40 randomly selected GENIA training texts. In the rest of this paper, all performances of using *40 GENIA training texts* are the average scores over 5 runs, each of which uses a different set of 40 texts.

In the previous section, we have presented the domain adaptation techniques, the active learning algorithm, as well as the target domain instance weighting scheme. In the rest of this section, we present the experimental results to show how domain adaptation, active learning, and target domain instance weighting help coreference resolution in a new domain. We use *Augment*, *IW*, and *IP* to denote the three domain adaptation techniques: AUGMENT, INSTANCE WEIGHTING, and INSTANCE PRUNING, respectively. For a further comparison, we explore another baseline method, which is simply a concatenation of the source and target domain data together, called *Combine* in the rest of this paper. In all the experiments with active learning, we run 100 iterations, which result in the selection of 1,000 target domain instances.

The first experiment is to measure the effectiveness of target domain instance weighting. We fix on the use of uncertainty-based active learning, and compare weighting and without weighting of target domain instances (denoted as *Weighted* and *Unweighted*). The learning curves are shown in Figure 2. For *Combine*, *Augment*, and *IP*, it can be seen that *Weighted* is a clear winner. As for *IW*, at the beginning of active learning, *Unweighted* outperforms *Weighted*, though it is unstable. At the end of 100 iterations, *Weighted* outperforms *Unweighted*.

Since *Weighted* outperforms *Unweighted*, we fix on the use of *Weighted* and explore the effectiveness of active learning. For comparison, we try another iterative process that randomly selects 10 instances in each iteration. We found that selection of instances using active learning achieved better performance than random selection in all cases. This is because random selection may select instances that the classifier has very high confidence in, which will not help in improving the classifier.

In the third experiment, we fix on the use of *Weighted* and *Uncertainty* since they perform the best, and evaluate the effect of different domain adaptation techniques. The learning curves are shown in Figure 3. It can be seen that *Augment* is the best performing system. For a closer look, we tabulate the results in Table 3, with the statistical significance levels indicated. Statistical significance tests were conducted following (Chinchor, 2011).



(a) Combine

(b) Augment

(c) IW

(d) IP

Figure 2: Learning curves of comparing target domain instances weighted vs. unweighted. All systems use uncertainty-based active learning.

| Iteration | 0 | 10 | 20 | 30 | 40 | 60 | 80 | 100 |
|---|---|---|---|---|---|---|---|---|
| Combine+Unweighted | 39.8 | 40.7 | 40.9 | 41.1 | 41.4 | 40.4 | 41.6 | 42.1 |
| Combine+Weighted | 39.8 | 40.9 | 44.0** | 44.8** | 45.2** | 48.0** | 47.7** | 47.6** |
| Augment+Weighted | 39.8 | **44.1****†† | **46.0****†† | **47.0****†† | **47.8****†† | **49.1****†† | **49.1****†† | **49.0****†† |
| IW+Weighted | 39.8 | 24.3 | 33.1 | 36.8 | 38.1 | 45.0** | 48.2**†† | 48.3**†† |
| IP+Weighted | 39.8 | 34.4 | 40.7 | 43.4** | 46.2**†† | 48.0** | 48.5**†† | 48.5**†† |

Table 3: MUC F-measures of different active learning settings on the GENIA test set. All systems use *Uncertainty*. Statistical significance is compared against *Combine+Unweighted*, where * and ** stand for $p < 0.05$ and $p < 0.01$, respectively, and compared against *Combine+Weighted*, where †and ††stand for $p < 0.05$ and $p < 0.01$, respectively.

## 6 Analysis

Using only the source domain training data, a coreference resolution system achieves an F-measure of 39.8% on the GENIA test set (the column of "Iteration 0" in Table 3). From Figure 3 and Table 3, we can see that in the first few iterations of active learning, domain adaptation does not perform as well as using only the source domain training data. This is because when there are very limited target domain data, the estimation of the target domain is unreliable. Dahlmeier and Ng (2010) reported similar findings though they did not use active learning. With more iterations, i.e., more target domain training data, domain adaptation is clearly superior. Among the three domain adaptation techniques, *Augment* is better than *IW* and *IP*. It not only achieves a higher F-measure, but also a faster speed to adapt to a new domain in active learning. Also, similar to (Dahlmeier and Ng, 2010), we find that *IP* is generally better than *IW*. All systems (except *IW*) with *Weighted* performs much better than *Combine+Unweighted*. This shows the effectiveness of target domain instance weighting. The average recall, precision, and F-measure of our best model, *Augment+Weighted*, after 100 iterations are 37.3%, 71.5%, and 49.0%, respectively. Compared to training with only the NPAPER training data, not only the F-measure, but also both the recall and precision are greatly improved (cf Table 2).

Among all the target domain instances that were selected in *Augment+Weighted*, the average dis-

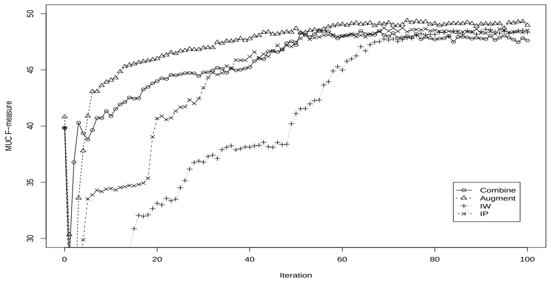


Figure 3: Learning curves of different domain adaptation methods. All systems use *Weighted* and *Uncertainty*.

tance of the two markables in an instance (measured in sentence) is 3.4 (averaged over the 5 runs), which means an annotator needs to read 4 sentences on average to annotate an instance.

We also investigate the difference of coreference resolution between the newswire domain and the biomedical domain, and the instances that were selected in active learning which represent this difference. One of the reasons that coreference resolution differs in the two domains is that scientific writing in biomedical texts frequently compares entities. For example,

> In Cushing's syndrome, the CR of GR was normal in spite of the fact that the CR of plasma cortisol was disturbed.

The two *CR*s refer to different entities and hence are not coreferential. However, a system trained on NPAPER predicts them as coreferential. In the newswire domain, comparisons are less likely, especially for named entities. For example, in the newswire domain, *London* in most cases is coreferential to other *London*s. However, in the biomedical domain, *DNA*s as in *DNA of human beings* and *DNA of monkeys* are different entities. A coreference resolution system trained on the newswire domain is unable to capture the difference between these two named entities, hence predicting them as coreferential. This also justifies the need for domain adaptation for coreference resolution. For the above sentence, after applying our method, the adapted coreference resolution system is able to predict the two *CR*s as non-coreferential.

Next, we show the effectiveness of our system using domain adaptation with active learning compared to a system trained with full coreference annotations. Averaged over 5 runs, a system trained on a single GENIA training text achieves an F-measure of 25.9%, which is significantly lower than that achieved by our method. With more GENIA training texts added, the F-measure increases. After 80 texts are used, the system trained on full annotations finally achieves an F-measure of 49.2%, which is 0.2% higher than *Augment+Weighted* after 100 iterations. However, after 100 iterations, only 1,000 target domain instances are annotated under our framework. Considering that one single text in the GENIA corpus contains an average of over 2,000 instances (cf Table 1), effectively we annotate only half of a text. Compared to the 80 training texts needed, this is a huge reduction. In order to achieve similar performance, we only need to annotate 1/160 or 0.63% of the complete set of training instances under our framework of domain adaptation with active learning.

Lastly, although in this paper we reported experimental results with the MUC evaluation metric, we also evaluated our approach with other evaluation metrics for coreference resolution, e.g., the B-CUBED metric, and obtained similar findings.

## 7 Conclusion

In this paper, we presented an approach using domain adaptation with active learning to adapt coreference resolution from the newswire domain to the biomedical domain. We explored the effect of domain adaptation, active learning, and target domain instance weighting for coreference resolution. Experimental results showed that domain adaptation with active learning and the target instance weighting scheme achieved a similar performance on MEDLINE abstracts but with a greatly reduced number of annotated training instances, compared to a system trained on full coreference annotations.

## Acknowledgments

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

## References

José Castaño, Jason Zhang, and James Pustejovsky. 2002. Anaphora resolution in biomedical literature. In *Proceedings of the International Symposium on Reference Resolution*.

Yee Seng Chan and Hwee Tou Ng. 2007. Domain adaptation with active learning for word sense disambiguation. In *Proceedings of the ACL2007*.

Nancy Chinchor. 2011. Statistical significance of MUC-6 results. In *Proceedings of the Sixth Message Understanding Conference*.

K. Bretonnel Cohen, Arrick Lanfranchi, William Corvey, William A. Baumgartner Jr., Christophe Roeder, Philip V. Ogren, Martha Palmer, and Lawrence Hunter. 2010. Annotation of all coreference in biomedical text: Guideline selection and adaptation. In *BioTxtM 2010*.

Daniel Dahlmeier and Hwee Tou Ng. 2010. Domain adaptation for semantic role labeling in the biomedical domain. *Bioinformatics*, 26(8):1098–1104.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the ACL2007*.

Caroline Gasperin and Ted Briscoe. 2008. Statistical anaphora resolution in biomedical texts. In *Proceedings of the COLING2008*.

Caroline Gasperin, Nikiforos Karamanis, and Ruth Seal. 2007. Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. In *Proceedings of the DAARC2007*.

Caroline Gasperin. 2009. Active learning for anaphora resolution. In *Proceedings of the NAACL-HLT2009 Workshop on Active Learning for Natural Language Processing*.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of the ACL2007*.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011a. Overview of BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*.

Youngjun Kim, Ellen Riloff, and Nathan Gilbert. 2011b. The taming of Reconcile as a biomedical coreference resolver. In *Proceedings of BioNLP Shared Task 2011 Workshop*.

Florian Laws, Florian Heimerl, and Hinrich Schütze. 2012. Active learning for coreference resolution. In *Proceedings of the NAACL2012*.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the SIGIR1994*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Timothy A. Miller, Dmitriy Dligach, and Guergana K. Savova. 2012. Active learning for coreference resolution. In *Proceedings of the BioNLP2012*.

Makoto Miwa, Paul Thompson, and Sophia Ananiadou. 2012. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, 28(13):1759–1765.

MUC-6. 1995. Coreference task definition (v2.3, 8 Sep 95). In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*.

MUC-7. 1998. Coreference task definition (v3.0, 13 Jul 97). In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the ACL2002*.

NIST. 2002. The ACE 2002 evaluation plan. ftp://jaguar.ncsl.nist.gov/ace/doc/ACE-EvalPlan-2002-v06.pdf.

Piyush Rai, Avishek Saha, Hal Daume, and Suresh Venkatasubramanian. 2010. Domain adaptation meets active learning. In *Proceedings of the NAACL-HLT2010 Workshop on Active Learning for Natural Language Processing*.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

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 ACL-IJCNLP2009*.

Min Tang, Xiaoqiang Luo, and Salim Roukos. 2002. Active learning for statistical natural language parsing. In *Proceedings of the ACL2002*.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the MUC-6*.

Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. 2004. Improving noun phrase coreference resolution by matching strings. In *Proceedings of the IJCNLP2004*.

Jian Bo Yang, Qi Mao, Qiao Liang Xiang, Ivor W. Tsang, Kian Ming A. Chai, and Hai Leong Chieu. 2012. Domain adaptation for coreference resolution: An adaptive ensemble approach. In *Proceedings of the EMNLP2012*.

Shanheng Zhao and Hwee Tou Ng. 2010. Maximum metric score training for coreference resolution. In *Proceedings of the COLING2010*.

Zhi Zhong, Hwee Tou Ng, and Yee Seng Chan. 2008. Word sense disambiguation using OntoNotes: An empirical study. In *Proceedings of the EMNLP2008*.