Profile Based Cross-Document Coreference Using Kernelized Fuzzy Relational Clustering

Jian Huang[†] **Sarah M. Taylor**[‡] **Jonathan L. Smith**[‡] **Konstantinos A. Fotiadis**[‡] **C. Lee Giles**[†]
[†]College of Information Sciences and Technology
Pennsylvania State University, University Park, PA 16802, USA
{jhuang, giles}@ist.psu.edu
[‡]Advanced Technology Office, Lockheed Martin IS&GS, Arlington, VA 22203, USA
{sarah.m.taylor, jonathan.l.smith, konstantinos.a.fotiadis}@lmco.com

Abstract

Coreferencing entities across documents in a large corpus enables advanced document understanding tasks such as question answering. This paper presents a novel cross document coreference approach that leverages the profiles of entities which are constructed by using information extraction tools and reconciled by using a within-document coreference module. We propose to match the profiles by using a learned ensemble distance function comprised of a suite of similarity specialists. We develop a kernelized soft relational clustering algorithm that makes use of the learned distance function to partition the entities into fuzzy sets of identities. We compare the kernelized clustering method with a popular fuzzy relation clustering algorithm (FRC) and show 5% improvement in coreference performance. Evaluation of our proposed methods on a large benchmark disambiguation collection shows that they compare favorably with the top runs in the SemEval evaluation.

1 Introduction

A named entity that represents a person, an organization or a geo-location may appear within and across documents in different forms. Cross document coreference (CDC) is the task of consolidating named entities that appear in multiple documents according to their real referents. CDC is a stepping stone for achieving intelligent information access to vast and heterogeneous text corpora, which includes advanced NLP techniques such as document summarization and question answering. A related and well studied task is within document coreference (WDC), which limits the scope of disambiguation to within the boundary of a document. When namesakes appear in an article, the author can explicitly help to disambiguate, using titles and suffixes (as in the example, "George Bush *Sr.* ... the *younger* Bush") besides other means. Cross document coreference, on the other hand, is a more challenging task because these linguistics cues and sentence structures no longer apply, given the wide variety of context and styles in different documents.

Cross document coreference research has recently become more popular due to the increasing interests in the web person search task (Artiles et al., 2007). Here, a search query for a person name is entered into a search engine and the desired outputs are documents clustered according to the identities of the entities in question. In our work, we propose to drill down to the sub-document mention level and construct an entity profile with the support of information extraction tools and reconciled with WDC methods. Hence our IE based approach has access to accurate information such as a person's mentions and geo-locations for disambiguation. Simple IR based CDC approaches (e.g. (Gooi and Allan, 2004)), on the other hand, may simply use all the terms and this can be detrimental to accuracy. For example, a biography of John F. Kennedy is likely to mention members of his family with related positions, besides references to other political figures. Even with careful word selection, these textual features can still confuse the disambiguation system about the true identity of the person.

We propose to handle the CDC task using a novel kernelized fuzzy relational clustering algorithm, which allows probabilistic cluster membership assignment. This not only addresses the intrinsic uncertainty nature of the CDC problem, but also yields additional performance improvement. We propose to use a specialist ensemble

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 414–422,
Suntec, Singapore, 2-7 August 2009.

learning approach to aggregate the diverse set of similarities in comparing attributes and relationships in entity profiles. Our approach is first fully described in Section 2. The effectiveness of the proposed method is demonstrated using real world benchmark test sets in Section 3. We review related work in cross document coreference and conclude in Section 5.

2 Methods

2.1 Document Level and Profile Based CDC

We make distinctions between document level and profile based cross document coreference. *Document level* CDC makes a simplifying assumption that a named entity (and its variants) in a document has one underlying real identity. The assumption is generally acceptable but may be violated when a document refers to namesakes at the same time (e.g. George W. Bush and George H. W. Bush referred to as George or President Bush). Furthermore, the context surrounding the person NE President Clinton can be counterproductive for disambiguating the NE Senator Clinton, with both entities likely to appear in a document at the same time. The simplified document level CDC has nevertheless been used in the WePS evaluation (Artiles et al., 2007), called the web people task.

In this work, we advocate *profile based* disambiguation that aims to leverage the advances in NLP techniques. Rather than treating a document as simply a bag of words, an information extraction tool first extracts NE's and their relationships. For the NE's of interest (i.e. persons in this work), a within-document coreference (WDC) module then links the entities deemed as referring to the same underlying identity into a WDC chain. This process includes both *anaphora resolution* (resolving 'He' and its antecedent 'President Clinton') and *entity tracking* (resolving 'Bill' and 'President Clinton'). Let $\mathcal{E} = \{\mathbf{e}_1, ..., \mathbf{e}_N\}$ denote the set of N chained entities (each corresponding to a WDC chain), provided as input to the CDC system. We intentionally do not distinguish which document each $\mathbf{e}_j$ belongs to, as profile based CDC can potentially rectify WDC errors by leveraging information across document boundaries. Each $\mathbf{e}_i$ is represented as a profile which contains the NE, its attributes and associated relationships, i.e. $\mathbf{e}_j =< e_{j,1}, ..., e_{j,L} >$ ($e_{j,l}$ can be a textual attribute or a pointer to another entity). The profile based CDC method generates a partition of $\mathcal{E}$, represented by a partition matrix U (where u_{ij} denotes the membership of an entity $\mathbf{e}_j$ to the i-th identity cluster). Therefore, the chained entities placed in a name cluster are deemed as coreferent.

Profile based CDC addresses a finer grained coreference problem in the mention level, enabled by the recent advances in IE and WDC techniques. In addition, profile based CDC facilitates user information consumption with structured information and short summary passages. Next, we focus on the relational clustering algorithm that lies at the core of the profile based CDC system. We then turn our attention to the specialist learning algorithm for the distance function used in clustering, capable of leveraging the available training data.

2.2 CDC Using Fuzzy Relational Clustering

2.2.1 Preliminaries

Traditionally, hard clustering algorithms (where $u_{ij} \in \{0, 1\}$) such as complete linkage hierarchical agglomerative clustering (Mann and Yarowsky, 2003) have been applied to the disambiguation problem. In this work, we propose to use fuzzy clustering methods (relaxing the membership condition to $u_{ij} \in [0, 1]$) as a better way of handling uncertainty in cross document coreference. First, consider the following motivating example,

Example. The named entity *President Bush* is extracted from the sentence "President Bush addressed the nation from the Oval Office Monday."

- Without additional cues, a hard clustering algorithm has to arbitrarily assign the mention "President Bush" to either the NE "George W. Bush" or "George H. W. Bush".
- A soft clustering algorithm, on the other hand, can assign equal probability to the two identities, indicating low entropy or high uncertainty in the solution. Additionally, the soft clustering algorithm can assign lower probability to the identity "Governor Jeb Bush", reflecting a less likely (though not impossible) coreference decision.

We first formalize the cross document coreference problem as a soft clustering problem, which minimizes the following objective function:

$$J_C(\mathcal{E}) = \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^m d^2(\mathbf{e}_j, \mathbf{v}_i) \qquad (1)$$

$$s.t. \quad \sum_{i=1}^{C} u_{ij} = 1 \text{ and } \sum_{j=1}^{N} u_{ij} > 0, u_{ij} \in [0, 1]$$

where $\mathbf{v}_i$ is a virtual (implicit) prototype of the i-th cluster ($\mathbf{e}_j, \mathbf{v}_i \in \mathcal{D}$) and m controls the fuzziness of the solution ($m > 1$; the solution approaches hard clustering as m approaches 1). We will further explain the generic distance function $d : \mathcal{D} \times \mathcal{D} \rightarrow \mathsf{R}$ in the next subsection. The goal of the optimization is to minimize the sum of deviations of patterns to the cluster prototypes. The clustering solution is a fuzzy partition $\mathcal{P}_\theta = \{\mathcal{C}_i\}$, where $\mathbf{e}_j \in \mathcal{C}_i$ if and only if $u_{ij} > \theta$.

We note from the outset that the optimization functional has the same form as the classical Fuzzy C-Means (FCM) algorithm (Bezdek, 1981), but major differences exist. FCM, as most object clustering algorithms, deals with *object data* represented in a vectorial form. In our case, the data is purely relational and only the mutual relationships between entities can be determined. To be exact, we can define the similarity/dissimilarity between a pair of attributes or relationships of the same type l between entities $\mathbf{e}_j$ and $\mathbf{e}_k$ as $s^{(l)}(\mathbf{e}_j, \mathbf{e}_k)$. For instance, the similarity between the occupations 'President' and 'Commander in Chief' can be computed using the JC semantic distance (Jiang and Conrath, 1997) with WordNet; the similarity of co-occurrence with other people can be measured by the Jaccard coefficient. In the next section, we propose to compute the relation strength $r(\cdot, \cdot)$ from the component similarities using aggregation weights learned from training data. Hence the N chained entities to be clustered can be represented as relational data using an $n \times n$ matrix $\mathbf{R}$, where $r_{j,k} = r(\mathbf{e}_j, \mathbf{e}_k)$. The Any Relation Clustering Algorithm (ARCA) (Corsini et al., 2005; Cimino et al., 2006) represents *relational data* as object data using their mutual relation strength and uses FCM for clustering. We adopt this approach to transform (*objectify*) a relational pattern $\mathbf{e}_j$ into an N dimensional vector $\mathbf{r}_j$ (i.e. the j-th row in the matrix $\mathbf{R}$) using a mapping $\Theta : \mathcal{D} \rightarrow \mathsf{R}^N$. In other words, each chained entity is represented as a vector of its relation strengths with all the entities. Fuzzy clusters can then be obtained by grouping closely related patterns using object clustering algorithm.

Furthermore, it is well known that FCM is a spherical clustering algorithm and thus is not generally applicable to relational data which may yield relational clusters of arbitrary and complicated shapes. Also, the distance in the transformed space may be non-Euclidean, rendering many clustering algorithms ineffective (many FCM extensions theoretically require the underlying distance to satisfy certain metric properties). In this work, we propose kernelized ARCA (called KARC) which uses a kernel-induced metric to handle the *objectified* relational data, as we introduce next.

2.2.2 Kernelized Fuzzy Clustering

Kernelization (Schölkopf and Smola, 2002) is a machine learning technique to transform patterns in the data space to a high-dimensional feature space so that the structure of the data can be more easily and adequately discovered. Specifically, a nonlinear transformation Φ maps data in R^N to H of possibly infinite dimensions (Hilbert space). The key idea is the *kernel trick* – without explicitly specifying Φ and H, the inner product in H can be computed by evaluating a kernel function K in the data space, i.e. $< \Phi(\mathbf{r}_i), \Phi(\mathbf{r}_j) >= K(\mathbf{r}_i, \mathbf{r}_j)$ (one of the most frequently used kernel functions is the Gaussian RBF kernel: $K(\mathbf{r}_j, \mathbf{r}_k) = \exp(-\lambda \|\mathbf{r}_j - \mathbf{r}_k\|^2)$). This technique has been successfully applied to SVMs to classify nonlinearly separable data (Vapnik, 1995). Kernelization preserves the simplicity in the formalism of the underlying clustering algorithm, meanwhile it yields highly nonlinear boundaries so that spherical clustering algorithms can apply (e.g. (Zhang and Chen, 2003) developed a kernelized object clustering algorithm based on FCM).

Let $\mathbf{w}_i$ denote the objectified virtual cluster $\mathbf{v}_i$, i.e. $\mathbf{w}_i = \Theta(\mathbf{v}_i)$. Using the kernel trick, the squared distance between $\Phi(\mathbf{r}_j)$ and $\Phi(\mathbf{w}_i)$ in the feature space H can be computed as:

$$\begin{aligned} & \|\Phi(\mathbf{r}_j) - \Phi(\mathbf{w}_i)\|_H^2 \qquad (2) \\ = \; & < \Phi(\mathbf{r}_j) - \Phi(\mathbf{w}_i), \Phi(\mathbf{r}_j) - \Phi(\mathbf{w}_i) > \\ = \; & < \Phi(\mathbf{r}_j), \Phi(\mathbf{r}_j) > -2 < \Phi(\mathbf{r}_j), \Phi(\mathbf{w}_i) > \\ & + < \Phi(\mathbf{w}_i), \Phi(\mathbf{w}_i) > \\ = \; & 2 - 2K(\mathbf{r}_j, \mathbf{w}_i) \qquad (3) \end{aligned}$$

assuming $K(\mathbf{r}, \mathbf{r}) = 1$. The KARC algorithm defines the generic distance d as $d^2(\mathbf{e}_j, \mathbf{v}_i) \stackrel{def}{=} \|\Phi(\mathbf{r}_j) - \Phi(\mathbf{w}_i)\|_H^2 = \|\Phi(\Theta(\mathbf{e}_j)) - \Phi(\Theta(\mathbf{v}_i))\|_H^2$ (we also use d_{ji}^2 as a notational shorthand).

Using Lagrange Multiplier as in FCM, the optimal solution for Equation (1) is:

$$u_{ij} = \begin{cases} \left[\sum_{h=1}^{C} \left(\frac{d_{ji}^2}{d_{jh}^2} \right)^{1/(m-1)} \right]^{-1} & , (d_{ji}^2 \neq 0) \\ 1 & , (d_{ji}^2 = 0) \end{cases} \qquad (4)$$

$$\Phi(\mathbf{w}_i) = \frac{\sum_{k=1}^{N} u_{ik}^m \Phi(\mathbf{r}_k)}{\sum_{k=1}^{N} u_{ik}^m} \quad (5)$$

Since Φ is an implicit mapping, Eq. (5) can not be explicitly evaluated. On the other hand, plugging Eq. (5) into Eq. (3), d_{ji}^2 can be explicitly represented by using the kernel matrix,

$$d_{ji}^2 = 2 - 2 \cdot \frac{\sum_{k=1}^{N} u_{ik}^m K(\mathbf{r}_j, \mathbf{r}_k)}{\sum_{k=1}^{N} u_{ik}^m} \quad (6)$$

With the derivation, the kernelized fuzzy clustering algorithm KARC works as follows. The chained entities $\mathcal{E}$ are first objectified into the relation strength matrix $\mathbf{R}$ using SEG, the details of which are described in the following section. The Gram matrix $\mathbf{K}$ is then computed based on the relation strength vectors using the kernel function. For a given number of clusters C, the initialization step is done by randomly picking C patterns as cluster centers, equivalently, C indices $\{n_1, .., n_C\}$ are randomly picked from $\{1, .., N\}$. D^0 is initialized by setting $d_{ji}^2 = 2 - 2K(\mathbf{r}_j, \mathbf{r}_{n_i})$. KARC alternately updates the membership matrix U and the kernel distance matrix D until convergence or running more than $maxIter$ iterations (Algorithm 1). Finally, the soft partition is generated based on the membership matrix U, which is the desired cross document coreference result.

Algorithm 1 KARC Alternating Optimization

Input: Gram matrix $\mathbf{K}$; #Clusters C; threshold θ
  initialize D^0
  $t \leftarrow 0$
  repeat
    $t \leftarrow t + 1$
    // 1– Update membership matrix U^t:
    $u_{ij} = \frac{(d_{ji}^2)^{-\frac{1}{m-1}}}{\sum_{h=1}^{C} (d_{jh}^2)^{-\frac{1}{m-1}}}$
    // 2– Update kernel distance matrix D^t:
    $d_{ji}^2 = 2 - 2 \cdot \frac{\sum_{k=1}^{N} u_{ik}^m K_{jk}}{\sum_{k=1}^{N} u_{ik}^m}$
  until $(t > maxIter)$ or
    $(t > 1$ and $|U^t - U^{t-1}| < \epsilon)$
  $\mathcal{P}_\theta \leftarrow$ `Generate_soft_partition`(U^t, θ)
Output: Fuzzy partition $\mathcal{P}_\theta$

2.2.3 Cluster Validation

In the CDC setting, the number of true underlying identities may vary depending on the entities' level of ambiguity (e.g. name frequency). Selecting the optimal number of clusters is in general a hard research question in clustering[1]. We adopt the Xie-Beni Index (XBI) (Xie and Beni, 1991) as in ARCA, which is one of the most popular cluster validities for fuzzy clustering algorithms. Xie-Beni Index (XBI) measures the goodness of clustering using the ratio of the intra-cluster variation and the inter-cluster separation. We measure the kernelized XBI (KXBI) in the feature space as,

$$\text{KXBI} = \frac{\sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^m \|\Phi(\mathbf{r}_j) - \Phi(\mathbf{w}_i)\|_H^2}{N \cdot \min_{1 \le i < j \le C} \|\Phi(\mathbf{w}_i) - \Phi(\mathbf{w}_j)\|_H^2}$$

where the nominator is readily computed using D and the inter-cluster separation in the denominator can be evaluated using the similar kernel trick above (details omitted). Note that KXBI is only defined for $C > 1$. Thus we pick the C that corresponds to the first minimum of KXBI, and then compare its objective function value J_C with the cluster variance (J_1 for $C = 1$). The optimal C is chosen from the minimum of the two[2].

2.3 Specialist Ensemble Learning of Relation Strengths between Entities

One remaining element in the overall CDC approach is how the relation strength $r_{j,k}$ between two entities is computed. In (Cohen et al., 2003), a binary SVM model is trained and its confidence in predicting the non-coreferent class is used as the distance metric. In our case of using information extraction results for disambiguation, however, only some of the similarity features are present based on the available relationships in two profiles. In this work, we propose to treat each similarity function as a *specialist* that specializes in computing the similarity of a particular type of relationship. Indeed, the similarity function between a pair of attributes or relationships may in itself be a sophisticated component algorithm. We utilize the specialist ensemble learning framework (Freund et al., 1997) to combine these component

[1] In particular, clustering algorithms that regularize the optimization with cluster size are not applicable in our case.

[2] In practice, the entities to be disambiguated tend to be dominated by several major identities. Hence performance generally does not vary much in the range of large C values.

similarities into the relation strength for clustering. Here, a specialist is *awakened* for prediction only when the same type of relationships are present in both chained entities. A specialist can choose not to make a prediction if it is not confident enough for an instance. These aspects contrast with the traditional *insomniac* ensemble learning methods, where each component learner is always available for prediction (Freund et al., 1997). Also, specialists have different weights (in addition to their prediction) on the final relation strength, e.g. a match in a family relationship is considered more important than in a co-occurrence relationship.

Algorithm 2 SEG (Freund et al., 1997)

Input: Initial weight distribution $\mathbf{p^1}$;
learning rate $\eta > 0$; training set $\{< \mathbf{s}^t, y^t >\}$

1: **for** t=1 to T **do**
2: Predict using:

$$\tilde{y}^t = \frac{\sum_{i \in E^t} p_i^t s_i^t}{\sum_{i \in E^t} p_i^t} \quad (7)$$

3: Observe the true label y^t and incur square loss $L(\tilde{y}^t, y^t) = (\tilde{y}^t - y^t)^2$
4: Update weight distribution: for $i \in E_t$

$$p_i^{t+1} = \frac{p_i^t e^{-2\eta x_i^t (\tilde{y}^t - y^t)}}{\sum_{j \in E^t} p_j^t e^{-2\eta x_i^t (\tilde{y}^t - y^t)}} \cdot \sum_{j \in E_t} p_j^t \quad (8)$$

Otherwise: $p_i^{t+1} = p_i^t$
5: **end for**

Output: Model $\mathbf{p}$

The ensemble relation strength model is learned as follows. Given training data, the set of chained entities $\mathcal{E}_{train}$ is extracted as described earlier. For a pair of entities $\mathbf{e}_j$ and $\mathbf{e}_k$, a similarity vector $\mathbf{s}$ is computed using the component similarity functions for the respective attributes and relationships, and the true label is defined as $y = I\{\mathbf{e}_j$ and $\mathbf{e}_k$ are coreferent$\}$. The instances are subsampled to yield a balanced pairwise training set $\{< \mathbf{s}^t, y^t >\}$. We adopt the Specialist Exponentiated Gradient (SEG) (Freund et al., 1997) algorithm to learn the mixing weights of the specialists' prediction (Algorithm 2) in an online manner. In each training iteration, an instance $< \mathbf{s}^t, y^t >$ is presented to the learner (with E^t denoting the set of indices of awake specialists in $\mathbf{s}^t$). The SEG algorithm first predicts the value $\tilde{y}^t$ based on the awake specialists' decisions. The true value y^t is then revealed and the learner incurs a square loss between the predicted and the true values. The current weight distribution $\mathbf{p}$ is updated to minimize square loss: awake specialists are promoted or demoted in their weights according to the difference between the predicted and the true value. The learning iterations can run a few passes till convergence, and the model is learned in linear time with respect to T and is thus very efficient. In prediction time, let $E^{(jk)}$ denote the set of active specialists for the pair of entities $\mathbf{e}_j$ and $\mathbf{e}_k$, and $\mathbf{s}^{(jk)}$ denote the computed similarity vector. The predicted relation strength $r_{j,k}$ is,

$$r_{j,k} = \frac{\sum_{i \in E^{(jk)}} p_i s_i^{(jk)}}{\sum_{i \in E^{(jk)}} p_i} \quad (9)$$

2.4 Remarks

Before we conclude this section, we make several comments on using fuzzy clustering for cross document coreference. First, instead of conducting CDC for all entities concurrently (which can be computationally intensive with a large corpus), chained entities are first distributed into non-overlapping blocks. Clustering is performed for each block which is a drastically smaller problem space, while entities from different blocks are unlikely to be coreferent. Our CDC system uses phonetic blocking on the full name, so that name variations arising from translation, transliteration and abbreviation can be accommodated. Additional link constraints checking is also implemented to improve scalability though these are not the main focus of the paper.

There are several additional benefits in using a fuzzy clustering method besides the capability of probabilistic membership assignments in the CDC solution. In the clustered web search context, splitting a true identity into two clusters is perceived as a more severe error than putting irrelevant records in a cluster, as it is more difficult for the user to collect records in different clusters (to reconstruct the real underlying identity) than to prune away noisy records. While there is no universal way to handle this with hard clustering, soft clustering algorithms can more easily avoid the false negatives by allowing records to probabilistically appear in different clusters (subject to the sum of 1) using a more lenient threshold. Also, while there is no real prototypical elements in relational clustering, soft relational clustering

methods can naturally rank the profiles within a cluster according to their membership levels, which is an additional advantage for enhancing user consumption of the disambiguation results.

3 Experiments

In this section, we first formally define the evaluation metrics, followed by the introduction to the benchmark test sets and the system's performance.

3.1 Evaluation Metrics

We benchmarked our method using the standard purity and inverse purity clustering metrics as in the WePS evaluation. Let a set of clusters $\mathcal{P} = \{\mathcal{C}_i\}$ denote the system's partition as aforementioned and a set of categories $\mathcal{Q} = \{\mathcal{D}_j\}$ be the gold standard. The precision of a cluster $\mathcal{C}_i$ with respect to a category $\mathcal{D}_j$ is defined as,

$$\text{Precision}(\mathcal{C}_i, \mathcal{D}_j) = \frac{|\mathcal{C}_i \cap \mathcal{D}_j|}{|\mathcal{C}_i|}$$

Purity is in turn defined as the weighted average of the maximum precision achieved by the clusters on one of the categories,

$$\text{Purity}(\mathcal{P}, \mathcal{Q}) = \sum_{i=1}^{C} \frac{|\mathcal{C}_i|}{n} \max_j \text{Precision}(\mathcal{C}_i, \mathcal{D}_j)$$

where $n = \sum |\mathcal{C}_i|$. Hence purity penalizes putting noise chained entities in a cluster. Trivially, the maximum purity (i.e. 1) can be achieved by making one cluster per chained entity (referred to as the one-in-one baseline). Reversing the role of clusters and categories, $\text{Inverse\_purity}(\mathcal{P}, \mathcal{Q}) \stackrel{def}{=} \text{Purity}(\mathcal{Q}, \mathcal{P})$. Inverse Purity penalizes splitting chained entities belonging to the same category into different clusters. The maximum inverse purity can be similarly achieved by putting all entities into one cluster (all-in-one baseline).

Purity and inverse purity are similar to the precision and recall measures commonly used in IR. The F score, $F = 1/(\alpha\frac{1}{Purity} + (1-\alpha)\frac{1}{InversePurity})$, is used in performance evaluation. $\alpha = 0.2$ is used to give more weight to inverse purity, with the justification for the web person search mentioned earlier.

3.2 Dataset

We evaluate our methods using the benchmark test collection from the ACL SemEval-2007 web person search task (WePS) (Artiles et al., 2007). The test collection consists of three sets of 10 different names, sampled from ambiguous names from English Wikipedia (famous people), participants of the ACL 2006 conference (computer scientists) and common names from the US Census data, respectively. For each name, the top 100 documents retrieved from the Yahoo! Search API were annotated, yielding on average 45 real world identities per set and about 3k documents in total.

As we note in the beginning of Section 2, the human markup for the entities corresponding to the search queries is on the document level. The profile-based CDC approach, however, is to merge the mention-level entities. In our evaluation, we adopt the document label (and the person search query) to annotate the entity profiles that corresponds to the person name search query. Despite the difference, the results of the one-in-one and all-in-one baselines are almost identical to those reported in the WePS evaluation ($F = 0.52, 0.58$ respectively). Hence the performance reported here is comparable to the official evaluation results (Artiles et al., 2007).

3.3 Information Extraction and Similarities

We use an information extraction tool AeroText (Taylor, 2004) to construct the entity profiles. AeroText extracts two types of information for an entity. First, the attribute information about the person named entity includes first/middle/last names, gender, mention, etc. In addition, AeroText extracts relationship information between named entities, such as Family, List, Employment, Ownership, Citizen-Resident-Religion-Ethnicity and so on, as specified in the ACE evaluation. AeroText resolves the references of entities within a document and produces the entity profiles, used as input to the CDC system. Note that alternative IE or WDC tools, as well as additional attributes or relationships, can be readily used in the CDC methods we proposed.

A suite of similarity functions is designed to determine if the attributes relationships in a pair of entity profiles match or not:

Text similarity. To decide whether two names in the co-occurrence or family relationship match, we use the SoftTFIDF measure (Cohen et al., 2003), which is a hybrid matching scheme that combines the token-based TFIDF with the Jaro-Winkler string distance metric. This permits inexact matching of named entities due to name

variations, typos, etc.

Semantic similarity. Text or syntactic similarity is not always sufficient for matching relationships. WordNet and the information theoretic semantic distance (Jiang and Conrath, 1997) are used to measure the semantic similarity between concepts in relationships such as mention, employment, ownership, etc.

Other rule-based similarity. Several other cases require special treatment. For example, the employment relationships of *Senator* and *D-N.Y.* should match based on domain knowledge. Also, we design dictionary-based similarity functions to handle nicknames (Bill and William), acronyms (COLING for International Conference on Computational Linguistics), and geo-locations.

3.4 Evaluation Results

From the WePS training data, we generated a training set of around 32k pairwise instances as previously stated in Section 2.3. We then used the SEG algorithm to learn the weight distribution model. We tuned the parameters in the KARC algorithm using the training set with discrete grid search and chose $m = 1.6$ and $\theta = 0.3$. The RBF kernel (Gaussian) is used with $\gamma = 0.015$.

Table 1: Cross document coreference performance (I. Purity denotes inverse purity).

Method	Purity	I. Purity	F
KARC-S	0.657	0.795	0.740
KARC-H	0.662	0.762	0.710
FRC	0.484	0.840	0.697
One-in-one	1.000	0.482	0.524
All-in-one	0.279	1.000	0.571

The macro-averaged cross document coreference on the WePS test sets are reported in Table 1. The F score of our CDC system (KARC-S) is 0.740, comparable to the test results of the first tier systems in the official evaluation. The two baselines are also included. Since different feature sets, NLP tools, etc are used in different benchmarked systems, we are also interested in comparing the proposed algorithm with different soft relational clustering variants. First, we 'harden' the fuzzy partition produced by KARC by allowing an entity to appear in the cluster with highest membership value (KARC-H). Purity improves because of the removal of noise entities, though at the sacrifice of inverse purity and the F score deteriorates. We also implement a popular fuzzy relational clustering algorithm called FRC (Dave and Sen, 2002), whose optimization functional directly minimizes with respect to the relation matrix. With the same feature sets and distance function, KARC-S outperforms FRC in F score by about 5%. Because the test set is very ambiguous (on average only two documents per real world entity), the baselines have relatively high F score as observed in the WePS evaluation (Artiles et al., 2007). Table 2 further analyzes KARC-S's result on the three subsets Wikipedia, ACL06 and US Census. The F score is higher in the less ambiguous (the average number of identities) dataset and lower in the more ambiguous one, with a spread of 6%.

Table 2: Cross document coreference performance on subsets (I. Purity denotes inverse purity).

Test set	Identity	Purity	I. Purity	F
Wikipedia	56.5	0.666	0.752	0.717
ACL-06	31.0	0.783	0.771	0.773
US Census	50.3	0.554	0.889	0.754

We study how the cross document coreference performance changes as we vary the fuzziness in the solution (controlled by m). In Figure 1, as m increases from 1.4 to 1.9, purity improves by 10% to 0.67, which indicates that more correct coreference decisions (true positives) can be made in a softer configuration. The complimentary is true for inverse purity, though to a lesser extent. In this case, more false negatives, corresponding to the entities of different coreferents incorrectly

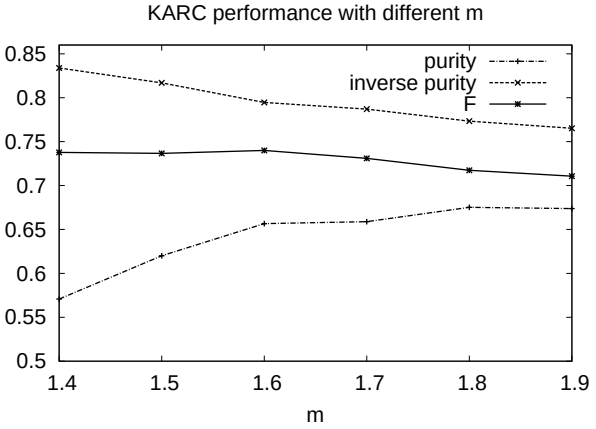

Figure 1: Purity, inverse purity and F score with different fuzzifiers m.

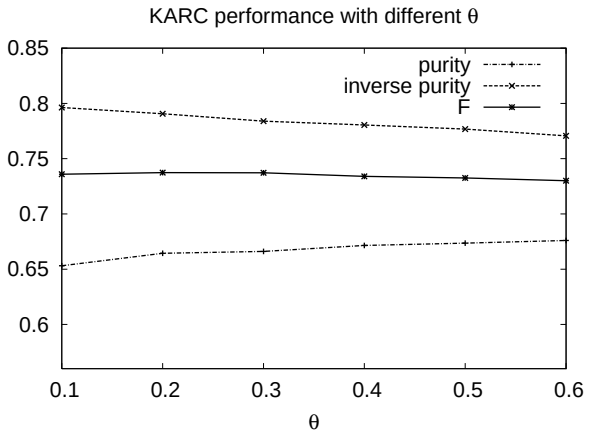

Figure 2: CDC performance with different θ.

linked, are made in a softer partition. The F score peaks at 0.74 ($m = 1.6$) and then slightly decreases, as the gain in purity is outweighed by the loss in inverse purity.

Figure 2 evaluates the impact of the different settings of θ (the threshold of including a chained entity in the fuzzy cluster) on the coreference performance. We observe that as we increase θ, purity improves indicating less 'noise' entities are included in the solution. On the other hand, inverse purity decreases meaning more coreferent entities are not linked due to the stricter threshold. Overall, the changes in the two metrics offset each other and the F score is relatively stable across a broad range of θ settings.

4 Related Work

The original work in (Bagga and Baldwin, 1998) proposed a CDC system by first performing WDC and then disambiguating based on the summary sentences of the chains. This is similar to ours in that mentions rather than documents are clustered, leveraging the advances in state-of-the-art WDC methods developed in NLP, e.g. (Ng and Cardie, 2001; Yang et al., 2008). On the other hand, our work goes beyond the simple bag-of-word features and vector space model in (Bagga and Baldwin, 1998; Gooi and Allan, 2004) with IE results. (Wan et al., 2005) describes a person resolution system WebHawk that clusters web pages using some extracted personal information including person name, title, organization, email and phone number, besides lexical features. (Mann and Yarowsky, 2003) extracts biographical information, which is relatively scarce in web data, for disambiguation. With the support of state-of-the-art information extraction tools, the profiles of entities in this work covers a broader range of relational information. (Niu et al., 2004) also leveraged IE support, but their approach was evaluated on a small artificial corpus. Also, the pairwise distance model is *insomniac* (i.e. all similarity specialists are *awake* for prediction) and our work extends this with a specialist learning framework.

Prior work has largely relied on using hierarchical clustering methods for CDC, with the threshold for stopping the merging set using the training data, e.g. (Mann and Yarowsky, 2003; Chen and Martin, 2007; Baron and Freedman, 2008). The fuzzy relational clustering method proposed in this paper we believe better addresses the uncertainty aspect of the CDC problem.

There are also orthogonal research directions for the CDC problem. (Li et al., 2004) solved the CDC problem by adopting a probabilistic view on how documents are generated and how names are sprinkled into them. (Bunescu and Pasca, 2006) showed that external information from Wikipedia can improve the disambiguation performance.

5 Conclusions

We have presented a profile-based Cross Document Coreference (CDC) approach based on a novel fuzzy relational clustering algorithm KARC. In contrast to traditional hard clustering methods, KARC produces fuzzy sets of identities which better reflect the intrinsic uncertainty of the CDC problem. Kernelization, as used in KARC, enables the optimization of clustering that is spherical in nature to apply to relational data that tend to have complicated shapes. KARC partitions named entities based on their profiles constructed by an information extraction tool. To match the profiles, a specialist ensemble algorithm predicts the pairwise distance by aggregating the similarities of the attributes and relationships in the profiles. We evaluated the proposed methods with experiments on a large benchmark collection and demonstrate that the proposed methods compare favorably with the top runs in the SemEval evaluation.

The focus of this work is on the novel learning and clustering methods for coreference. Future research directions include developing rich feature sets and using corpus level or external information. We believe that such efforts can further improve cross document coreference performance.

References

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. 2007. The SemEval-2007 WePS evaluation: Establishing a benchmark for the web people search task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 64–69.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of 36th International Conference On Computational Linguistics (ACL) and 17th international conference on Computational linguistics (COLING)*, pages 79–85.

Alex Baron and Marjorie Freedman. 2008. Who is who and what is what: Experiments in cross-document co-reference. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 274–283.

J. C. Bezdek. 1981. *Pattern Recognition with Fuzzy Objective Function Algoritms*. Plenum Press, NY.

Razvan Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 9–16.

Ying Chen and James Martin. 2007. Towards robust unsupervised personal name disambiguation. In *Proc. of 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Mario G. C. A. Cimino, Beatrice Lazzerini, and Francesco Marcelloni. 2006. A novel approach to fuzzy clustering based on a dissimilarity relation extracted from data using a TS system. *Pattern Recognition*, 39(11):2077–2091.

William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. 2003. A comparison of string distance metrics for name-matching tasks. In *Proceedings of IJCAI Workshop on Information Integration on the Web*.

Paolo Corsini, Beatrice Lazzerini, and Francesco Marcelloni. 2005. A new fuzzy relational clustering algorithm based on the fuzzy c-means algorithm. *Soft Computing*, 9(6):439 – 447.

Rajesh N. Dave and Sumit Sen. 2002. Robust fuzzy clustering of relational data. *IEEE Transactions on Fuzzy Systems*, 10(6):713–727.

Yoav Freund, Robert E. Schapire, Yoram Singer, and Manfred K. Warmuth. 1997. Using and combining predictors that specialize. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing (STOC)*, pages 334–343.

Chung H. Gooi and James Allan. 2004. Cross-document coreference on a large scale corpus. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 9–16.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics*.

Xin Li, Paul Morie, and Dan Roth. 2004. Robust reading: Identification and tracing of ambiguous names. In *Proceedings of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 17–24.

Gideon S. Mann and David Yarowsky. 2003. Unsupervised personal name disambiguation. In *Conference on Computational Natural Language Learning (CoNLL)*, pages 33–40.

Vincent Ng and Claire Cardie. 2001. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 104–111.

Cheng Niu, Wei Li, and Rohini K. Srihari. 2004. Weakly supervised learning for cross-document person name disambiguation supported by information extraction. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL)*, pages 597–604.

Bernhard Schölkopf and Alex Smola. 2002. *Learning with Kernels*. MIT Press, Cambridge, MA.

Sarah M. Taylor. 2004. Information extraction tools: Deciphering human language. *IT Professional*, 6(6):28 – 34.

Vladimir Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag New York.

Xiaojun Wan, Jianfeng Gao, Mu Li, and Binggong Ding. 2005. Person resolution in person search results: WebHawk. In *Proceedings of the 14th ACM international conference on Information and knowledge management (CIKM)*, pages 163–170.

Xuanli Lisa Xie and Gerardo Beni. 1991. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841 – 847.

Xiaofeng Yang, Jian Su, Jun Lang, Chew L. Tan, Ting Liu, and Sheng Li. 2008. An entity-mention model for coreference resolution with inductive logic programming. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 843–851.

Dao-Qiang Zhang and Song-Can Chen. 2003. Clustering incomplete data using kernel-based fuzzy c-means algorithm. *Neural Processing Letters*, 18(3):155 – 162.