Enhancing Cross Document Coreference of Web Documents with Context Similarity and Very Large Scale Text Categorization

Jian Huang
Information Sciences and Technology
Pennsylvania State University
jhuang@ist.psu.edu

Pucktada Treeratpituk
Information Sciences and Technology
Pennsylvania State University
pxt162@ist.psu.edu

Sarah M. Taylor
Lockheed Martin IS&GS
sarah.m.taylor@lmco.com

C. Lee Giles
Information Sciences and Technology
Pennsylvania State University
giles@ist.psu.edu

Abstract

Cross Document Coreference (CDC) is the task of constructing the coreference chain for mentions of a person across a set of documents. This work offers a holistic view of using document-level categories, sub-document level context and extracted entities and relations for the CDC task. We train a categorization component with an efficient flat algorithm using thousands of ODP categories and over a million web documents. We propose to use ranked categories as coreference information, particularly suitable for web documents that are widely different in style and content. An ensemble composite coreference function, amenable to inactive features, combines these three levels of evidence for disambiguation.

A thorough feature importance study is conducted to analyze how these three components contribute to the coreference results. The overall solution is evaluated using the WePS benchmark data and demonstrate superior performance.

1 Introduction

Cross Document Coreference (CDC) is the task to determine whether Named Entities (NE) from different documents refer to the same underlying identity. CDC enables a range of advanced NLP applications such as automated text summarization and question answering (e.g. list-type questions). CDC has mainly been developed from two perspectives.

First, in the Message Understanding Conference (MUC-6), CDC was viewed as an advanced task performed based on a set of Information Extraction (IE) artifacts. IE has been one of the central topics in NLP since the 1970s and gained much success in transforming natural language text to structured text. IE on the Web, however, is inherently very challenging. For one, the Web is comprised of such heterogenous content that IE systems, many of which are developed on tidy and domain-specific corpora, may achieve relatively limited coverage. Also, the content of web documents may not even be in the natural language form. Hence, though IE based features are quite precise, it is rather difficult to achieve good coverage that's necessary to disambiguate person entities on the Web.

Recently, there is significant research interest in a related task called Web Person Search (WePS) (Artiles et al., 2007), which seeks to determine whether two documents refer to the same person given a person name search query. Many systems employed the simple vector space model and word co-occurrence features for this task. Though more robust with better coverage, these methods are more susceptible to irrelevant words with regard to the entity of interest.

Rather than relying solely on IE based or word co-occurrence features, this work adopts a holistic view of the different types of features useful for cross document coreference. Specifically, the main features of our proposed CDC approach are:

- The proposed approach covers the entire spectrum of document level, sub-document context level and entity/relation level disambiguation evidence. In particular, we propose to use document categories as robust document level evidence. This comprehensive design naturally combines state-of-the-art categorization, information extraction and IE-driven IR methods and compensates the limitation of each of them.
- The features used in this work are domain independent and thus are particularly suitable for coreferencing web documents.
- The composite pairwise coreference function in this work can readily incorporate a set of heterogenous features that are not always active or are in different ranges, making it easily extensible to additional features. Moreover, we thoroughly study the contribution of each component and its features to gain insight on improving cross document coreference performance.

In this work, three components specialize in generating the aforementioned three levels of features as coreference decisions. Thus we refer to them as *experts*. After reviewing prior work on CDC, we describe the methods of each of these components in detail and present empirical results where appropriate. We then show how these components (and its features) are aggregated to predict pairwise coreference using an ensemble method. We evaluate the contribution of each component and the overall CDC results on a benchmark dataset. Finally, we conclude and discuss future work.

2 Related Work

Compared to the traditional (within-document) coreference resolution problem, cross document coreference is a much harder problem due to the divergence of contents and the lack of consistent discourse information across documents.

(Bagga and Baldwin, 1998b) presented one of the first CDC systems, which relied solely on the contextual words of the named entities. (Gooi and Allan, 2004) used a 55-word window as the context without significant accuracy penalty. As these approaches only considered word co-occurrence, they were more susceptible to genre differences. Recent CDC work has sought Information Extraction (IE) support. Extracted NEs and relationships were considered in (Niu et al., 2004) for improved CDC performance.

Many of these earlier CDC methods were evaluated on small and tidy news articles. CDC for Web documents is even more challenging. (Wan et al., 2005) proposed a web person resolution system called WebHawk, which extracted several attributes such as title, organization, email and phone number using patterns. These features however only covered small amount of disambiguation evidence and certain types of web pages (such as personal home pages). The more recent Web Person Search (WePS) task (Artiles et al., 2007) has created a benchmark dataset which is also used in this work. Different from CDC which aims to resolve mention level NEs, WePS distinguishes *documents* retrieved by a name search query according to the underlying identity. The top-performing system (Chen and Martin, 2007) in this task extracted phrasal contextual and document-level entities as rich features for coreference. Similar IR features are also used by other WePS systems as they are more robust to the variety of web pages (Artiles et al., 2007).

Instead of focusing on local information, (Li et al., 2004) proposed a generative model of entity co-occurrence to capture global document level information. However, inference in generative models is expensive for large scale web data. Our work instead considers document categories/topics that can be efficiently predicted and easily interpretable by users. Hand-tuned weights were used in (Baron and Freedman, 2008) and a linear classifier was used in (Li et al., 2004) to combine the extracted features. Our composite pairwise coreference function is based on an ensemble classifier and is more robust and capable of handling inactive features.

3 Text Categorization Aided CDC

Consider the following scenario for motivation. When a user searches for 'Michael Jordan', the official web page of the basketball player

‘Michael Jordan’[1] contains mostly his career statistics, whereas the homepage of ‘Michael I. Jordan’ the professor[2] contains his titles, contact information and advising students. Neither of these pages contain complete natural language sentences that most IE and NLP tools are designed to process. We propose to use document categories (trained from a very large scale and general purpose taxonomy, Open Directory Project (ODP)) as document level features for CDC. In this example, one can easily differentiate these namesakes by categorizing the former as ‘Top/Sports/Basketball/Professional’ and the latter as ‘Top/Computer/Artificial Intelligence/Machine Learning’. We first introduce the method to categorize Web documents; then we show how to combine these categories for coreferencing.

3.1 Very Large Scale Text Categorization

To handle the web CDC problem, the catagorization component needs to be able to categorize documents of widely different topics. The Open Directory Project (ODP), the largest and most comprehensive human edited directory of the Web[3], contains hundreds of thousands of categories labeled for 2 million Web pages. Leveraging this vast amount of web data and the large Web taxonomy has called for the development of very efficient text categorization methods. There is significant research interest in scaling up to categorize millions of pages to thousands of categories and beyond, called the many class classification setting (Madani and Huang, 2008). Flat classification methods (e.g. (Crammer et al., 2006; Madani and Huang, 2008)), which treat hierarchical categories as flat classes, have been very successful due to their superior scalability and simplicity compared to classical hierarchical one-against-rest categorization. Flat methods also achieve high accuracy that is on par with, or better than, the traditional counterparts.

We adopt a flat multiclass online classification algorithm Passive Aggressive (PA) (Crammer et al., 2006) to predict ranked categories for web documents. For a categorization problem with C categories, PA associates each category k with a weight vector $\mathbf{w}^k$, called its *prototype*. The degree of confidence for predicting category k with respect to an instance $\mathbf{x}$[4] (both in online training and testing) is determined by the similarity between the instance and the prototype — the inner product $\mathbf{w}^k \cdot \mathbf{x}$. PA predicts a ranked list of categories according to this confidence.

PA is a family of online and large-margin based classifiers. Given an instance $(\mathbf{x}_t, y_t)$ during online learning, the multiclass margin $marg$ in PA[5] is the difference between the score of the true category y_t and that of the highest ranked false positive category s, i.e.

$$marg = \mathbf{w}^{y_t} \cdot \mathbf{x_t} - \mathbf{w}^s \cdot \mathbf{x_t} \tag{1}$$

where $s = \arg\max_{s \neq y_t} \mathbf{w}^s \cdot \mathbf{x_t}$.

A positive margin value indicates that the algorithm makes a correct prediction. One is however not only satisfied with a positive margin value, but also seeks to achieve a margin value of at least 1. When this is not satisfied, the online algorithm suffers a multiclass hinge loss:

$$\mathcal{L}_{mc}(\mathbf{w}; (\mathbf{x}_t, y_t)) = \begin{cases} 0 & marg \geq 1 \\ 1 - marg & \text{otherwise} \end{cases}$$

where $\mathbf{w} = (\mathbf{w}^1, .., \mathbf{w}^C)$ denotes the concatenation of the C prototypes (into a vector).

In an online learning step, the PA-II variant updates the category prototype with the solution of this constrained optimization problem,

$$\begin{aligned} \mathbf{w}_{t+1} &= \arg\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w} - \mathbf{w}_t\|^2 + \mathcal{A}\xi^2 & (2) \\ s.t. &\quad \mathcal{L}_{mc}(\mathbf{w}; (\mathbf{x}_t, y_t)) \leq \xi. & (3) \end{aligned}$$

Essentially, if the margin is met (also implying no misclassification), PA *passively* accepts the current solution. Otherwise, PA *aggressively* learns the new prototype which satisfies the loss constraint and stays as close to the one previously learned as possible. To cope with *label noise*, PA-II introduces a slack variable ξ in the optimization

[1] See *www.nba.com/playerfile/michael_jordan/index.html*

[2] See *www.eecs.berkeley.edu/∼jordan/*

[3] See *http://www.dmoz.org/about.html* for details.

[4] $\mathbf{x}$ is the vector representation of word frequencies of the corresponding document, L_2 normalized.

[5] For brevity of presentation, we consider the *single label* multiclass categorization setting.

for a gentler update, a technique previously employed to derive soft-margin classifiers (Vapnik, 1998). $\mathcal{A}$ is a parameter that controls the *aggressiveness* of the update.

The solution to the above optimization problem amounts to only changing the two prototypes violating the margin in the update step:

$$\mathbf{w}_{t+1}^{y_t} = \mathbf{w}_t^{y_t} + \tau \mathbf{x}_t \qquad \mathbf{w}_{t+1}^{s} = \mathbf{w}_t^{s} - \tau \mathbf{x}_t$$

where $\tau = \frac{\mathcal{L}_{mc}}{\|\mathbf{x}_t\|^2 + \frac{1}{2\mathcal{A}}}$.

To conclude, PA treats the hierarchy as flat categories for multiclass classification. It is similar to Multiclass Perceptron (Crammer and Singer, 2003) but only updates two vectors per iteration and thus is more efficient.

3.2 Categories as Coreference Evidence

Conceptually, the text categorization component can be viewed as a function that maps a document $\mathbf{d}$ to a ranked list of top K categories along with their respective confidence scores, i.e.

$$\phi(\mathbf{d}) = \{< c_1, s_1 >, .., < c_K, s_K >\}$$

We leverage these document categories to measure the pairwise similarity of any two documents, $\text{sim}(\phi(\mathbf{d}^u), \phi(\mathbf{d}^v))$, for entity disambiguation. Given a taxonomy $\mathcal{T}$, we first formally define the *affinity* between a category c and one of its ancestor category c' in $\mathcal{T}$ as:

$$\text{affinity}(c; c') = 1 - \frac{len(c, c')}{depth(\mathcal{T})}$$

where len is the length of the shortest path between the two categories and *depth(T)* denotes the depth of the taxonomy. In other words, affinity is the complementary of the normalized path length between c and its ancestor c'.

Using graph theory terminology, $\text{LCA}(c_1, c_2)$ denote the *lowest common ancestor* of two categories c_1 and c_2 in $\mathcal{T}$. Given two category lists, $\phi(\mathbf{d}^u) = \{< c_1^u, s_1^u >, .., < c_K^u, s_K^u >\}$ and $\phi(\mathbf{d}^v) = \{< c_1^v, s_1^v >, .., < c_K^v, s_K^v >\}$, we use the $LCA(c_i^u, c_j^v)$ of each category pair c_i^u and c_j^v as the basis to measure similarity. Formally, we transform $\phi(\mathbf{d}^u)$ to a $K \times K$ dimensional vector:

$$\vec{\mathbf{v}}(\mathbf{d}^u) = [\text{affinity}(c_i^u; LCA(c_i^u, c_j^v)) \cdot s_i^u]^T \quad (4)$$

where $i, j = 1..K$. In other words, we project $\phi(\mathbf{d}^u)$ into a vector in the space spanned by the LCAs of category pairs. Using the same bases, we can derive $\vec{\mathbf{v}}(\mathbf{d}^v)$ analogically.

With this transformation, $\phi(\mathbf{d}^u)$ and $\phi(\mathbf{d}^v)$ are expressed in the common bases, i.e. their LCAs. Therefore, the similarity between the top K categories of two documents can be measured by the inner product of these two vectors:

$$\text{sim}(\phi(\mathbf{d}^u), \phi(\mathbf{d}^v)) = \vec{\mathbf{v}}(\mathbf{d}^u) \cdot \vec{\mathbf{v}}(\mathbf{d}^v) \quad (5)$$

3.3 Empirical Studies

To handle the diverse topics of Web documents, we leverage the ODP data to train the many class categorization algorithm. The public ODP data contains 361,621 categories and links to over 2 million pages. We crawled the original web pages from these links, which yielded 1.9 million pages (50GB in size). The taxonomy was condensed to depth three[6] and then very rare categories (having less than 5 instances) were discarded. The data set is created with these categories and the vector representation of the term weights of the extracted raw text. This dataset has 1,889,683 instances and 4,891 categories in total. Finally, stratified 80-20 split was performed on this dataset, i.e. 1.5M pages for training and 377K pages for testing.

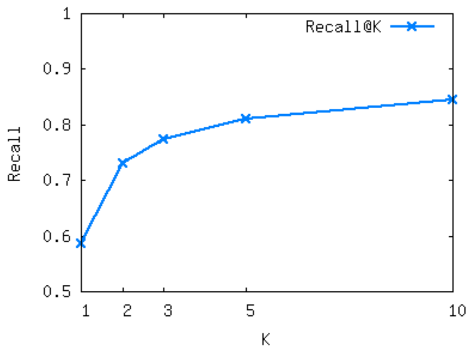

Figure 1: Categorization performance at different positions in the ODP test set.

As we view the taxonomy as a set of flat categories and we are interested in the top K categories, we use the recall at K metric for evaluation. Recall at K is defined as the percentage of instances having their true category ranked

[6] The original taxonomy has average depth 7, which is too deep for the coreference purpose in this work and many categories have too few instances for training.

among the top K slots in the category list. For a single label dataset (most ODP pages have one category) and $K = 1$, this is the accuracy metric in multiclass classification. Note that in the many class setting, recall at 1 is a very strict metric as no credit is given for predicting the parent, children or sibling categories; also, documents may have valid secondary topics not labeled by humans. Figure 1 shows recall at K in the test set. We observe that the algorithm is able to predict the category for 58.7% of the instances in the first rank and more than 77% in top three. There is only diminishing gains when we consider the categories further down the list. Hence we choose to use the similarity of the top 1 and top 3 categories (named `TC1` and `TC3`, respectively) and study their contributions for the CDC task.

3.4 Remarks

In this section, the entire document in the representation of its categories is used as a unit of analysis for CDC. Categorization based CDC works best with namesakes appearing in documents of relatively heterogenous topics, which is usually the case for web documents. Indeed, experienced web searchers would add terms such as 'baseball player' to the name search queries for more relevant results; Wikipedia also (manually) disambiguates namesakes by their professions. Categorization can also be adopted as a robust faceted search system for handling name search queries: users select the interested category/facet to efficiently disambiguate and filter out irrelevant results. The majority of web persons can be readily distinguished by the different underlying categories of the documents where they appear. For more homogeneous corpora or less benevolent cases, the next sections introduce two complementary CDC strategies.

4 Information Extraction for CDC

Consider the following two snippets retrieved with regard to the query 'George Bush':

[Snippet 1]: *"George W. Bush and Bill Clinton are trying to get Congress to allow Haiti to triple the number of exports ..."*

[Snippet 2]: *"George H. W. Bush succeeded Reagan as the 41st U.S. President."*

Using categories alone in this case is insufficient as both will be assigned similar categories such as 'Politics' or 'History/U.S.'. Also, it's not uncommon for these entities to co-occur in the same document and thus making them even more confounding. Properly disambiguating these two mentions requires the usage of local information: for instance, the extraction of full names, the detection of co-occurring NEs and contextual information. We introduce an IE system that extracts precise disambiguation evidence in this section and describe using the extraction context as additional information in the next section.

Our CDC system leverages a state-of-the-art commercial IE system AeroText (Taylor, 2004). The IE system employs manually created knowledge bases with statistically trained models to extract named entities, detect, classify and link relations between NEs. A summary of the most important IE-based features that we use are listed in Table 1. Based on the extracted attributes and relations, we further define their pairwise similarity used as coreference features. This ranges from simple compatibility checking for 'gender', textual soft matching for 'names', to sophisticated semantic matching for 'mentions' and 'locations' using WordNet. (Huang et al., 2009) provides more detailed discussions on the development of these IE based coreference features.

We note that several existing state-of-the-art IE systems are also capable of extracting these features. In particular, Named Entity Recognition (NER) which focuses on a small set of predefined categories of named entities (e.g. persons, organization, location) as well as the detection and tracking of preselected relations have achieved venerable empirical success in practice[7]. Also, within document coreference is a mature and well-studied technology in NLP (e.g. (Ng and Cardie, 2002)). Therefore, our CDC system can readily adopt alternative IE toolkits.

5 Context Matching

As mentioned earlier, achieving high extraction accuracy and coverage for diverse web documents

[7] The Automatic Content Extraction (ACE) evaluation and the Text Analysis Conference (TAC) also have IE-based entity tracking tasks that are relevant to this component.

is still a challenging and open research problem even for the state-of-the-art IE systems. We note that one of the natural outcomes from extraction is the context of the NE of interest, which covers the NE with its surrounding text. For a specific NE, our CDC system uses the context built from the sentences which form the NE's within document coreference chain. The context is then represented as a term vector whose terms are weighted by the TF-IDF weighing scheme. For a pair of NEs, the context matching component measures the cosine similarity of their context term vectors.

Essentially, this component alone is similar to the method presented in the seminal CDC work in (Bagga and Baldwin, 1998b). We however note that simply applying a predetermined threshold on the context similarity for CDC as in this earlier work is not sufficient. First, this method narrowly focuses on the local word occurrence and may miss the *big picture*, i.e. the correlation that exists in the global scope of a document. Also, mere word occurrence is incapable of accounting for the variation of word choices or placing special emphases on evidence such as co-occurring named entities, relations, etc. The categorization and IE components presented earlier in this work overcome these two pitfalls of the simple IR-based approach. We will further showcase the advantage of our comprehensive approach in section 7.2.

6 Composite Pairwise Coreference

In the previous sections, we describe the components to obtain document, sub-document and entity level disambiguation evidence in detail. In this section, we propose to use Random Forest (RF) to combine the experts components into one single composite pairwise similarity score. RF is an ensemble classifier, composed of a collection of randomized decision trees (Breiman, 2001). Each randomized tree is built on a different bootstrap sample of the training data. Randomness is also introduced into the tree construction process: the variable selection for each split is conducted not on the entire feature set, but from a small random subset of features. Gini index is used as the criteria in selecting the best split. Additionally, each tree is unpruned, to keep the prediction bias low. By aggregating many trees that are lowly-correlated (through bootstrap sampling and random variable selection), RF also reduces the prediction variance.

An ensemble method such as Random Forests is very suitable for the CDC task. First, the collection of randomized decision trees is analogous to a panel of different experts, where each makes its decision using different criteria and different features. Previously, RF has been used to aggregate various features in the author disambiguation task (Treeratpituk and Giles, 2009). One of the significant challenges in combining these different features in our CDC setting is that not all of them are always active. For instance, the IE tool may extract an employment relation for one entity and a list relation for another. Also, when the IE tool cannot infer the gender information or when the categorization component does not confidently predict the top K categories (e.g. all with low scores), it's desirable to not supply those features for coreferencing. The traditional technique to impute the missing values, e.g. by replacing them with the mean value, is not suitable in this case. In our work, we specify a special level 'NA' in the decision tree base learner. In our development set, this treatment improves pairwise coreference accuracy by more than 6%.

Figure 2 shows the convergence plot of the composite pairwise coreference function based on Random Forest[8]. We observe that the Out-Of-Bag

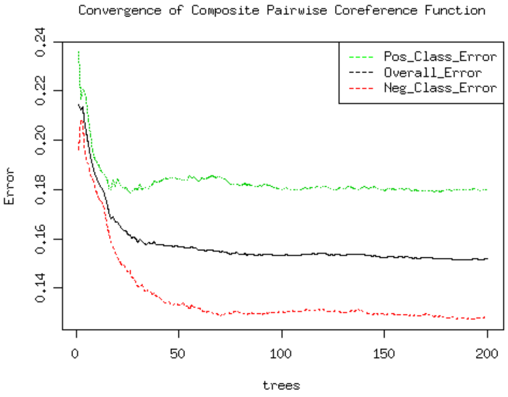

Figure 2: Convergence of OOB errors of the composite pairwise coreference function using the training portion of the WePS dataset.

[8] The R random forest (Liaw and Wiener, 2002) was used.

(OOB) errors [9] drastically decrease with the first 50 trees and then level off (without signs of overfitting). Thus we choose to use the model built with the first 100 trees for prediction. Overall, our model can achieve more than 85% accuracy for pairwise coreference prediction.

7 Experiments

We evaluate our CDC approach with the benchmark dataset from the ACL-2007 SemEval Web Person Search (WePS) evaluation campaign (Artiles et al., 2007). The WePS task is: given a name search query, cluster the search result *documents* according to the underlying referents. Compared to the CDC task which clusters *mention level* entities, a simplifying assumption is made in this task that each document refers to only one identity with respect to the query. The WePS dataset contains the training and test set. The training set contains the top 100 web search results of 49 names from the Web03 corpus (Mann and Yarowsky, 2003), Wikipedia and European Conference on Digital Library (ECDL) participants; the test data are comprised of the top 100 documents of 30 names from Wikipedia, US Census and ACL participants.

Table 1: Expert component and their feature sets.

Feature	Component	Description
TC1 TC3	Categorization	Sim. of the top 1 categories Sim. of the top 3 categories
CNTX	Context	Sim. of context
NAME MENT GEND	IE (attribute)	Sim. of full/first/last names Sim. of mentions Sim. of genders
EMP LIST LOC FAM	IE (relation)	Sim. of full/first/last names Sim. of co-occurring persons Sim. of locations Sim. of family members

7.1 Evaluation of Pairwise Coreference

We conduct a thorough study of the importance of the individual expert components and their features with the WePS training set. Table 1 shows the three components of the systems, their main features and descriptions.

The importance of these expert components and their features are illustrated in Figure 3. One of the most important features is CNTX, this confirms that the prior work on CDC (e.g. (Bagga and Baldwin, 1998b)) can achieve good results with the IE-driven context similarity feature (or its variation). The text categorization component also contributes very important features. In particular, TC3 is more significant than TC1 for reducing the Gini index because it recalls more correct categories. On the other hand, TC1 is slightly more important than TC3 for its contribution to accuracy, indicating TC1 is more precise (with less noise categories). For the IE component, attribute features NAME and MENT are the most useful. As aforementioned, the IE component may not always extract the relation features such as EMP, LIST, LOC and FAM, and hence they seemingly have limited effect on model learning (with relatively low reduction in Gini index). These relation features are however very accurate when extracted and are present for prediction. Therefore, they are strong disambiguation evidence and their removal would significantly hamper performance.

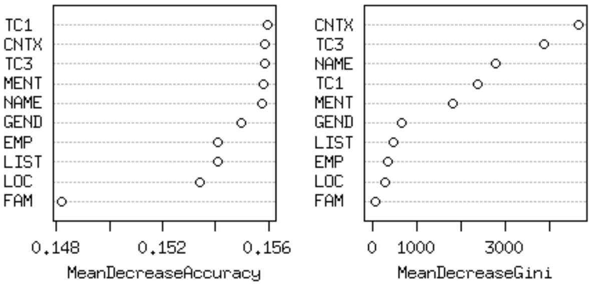

Figure 3: Importance of the expert components and their features found by Random Forest (note the small spread in MeanDecreaseAccuracy).

7.2 Evaluation for Web Person Search

Using the confidence of the pairwise coreference prediction as a distance metric, we adopt a density-based clustering method DBSCAN (Ester et al., 1996) as in (Huang et al., 2006)[10] to induce the person clusters. The final set of evaluation is based on these person clusters generated for the WePS test set.

Two sets of metrics are used to evaluate the overall system. First, we use the B-CUBED

[9] OOB error is an unbiased estimate of test error in RF (Breiman, 2001), computed as the average misclassification rates of each tree with samples not used for its construction.

[10] DBSCAN is a robust and scalable algorithm suitable for clustering relational data. In interest of space, we refer readers to (Ester et al., 1996) for the original algorithm.

Table 2: Cross document coreference performance (I. Pur. denotes inverse purity).

Method	Purity	I. Pur.	F	B-CUBED
CDC	0.812	0.796	**0.793**	**0.775**
CNTX	0.863	0.601	0.678	0.675
TC1+3	0.620	0.776	0.660	0.634
OIO	1.000	0.482	0.618	0.618
AIO	0.279	1.000	0.389	0.238

scores designed in (Bagga and Baldwin, 1998a) for evaluating cross document coreference performance. Second, we use the purity, inverse purity and their F score as in WePS (Artiles et al., 2007). Purity penalizes placing noise entities in a cluster, while inverse purity penalizes splitting coreferent entities into separate clusters.

Table 2 shows the performance of the macro-averaged cross document coreference performance on the WePS test sets. Note that though our evaluation is based on the mention level entities, the baselines One-In-One (OIO, placing each entity in a separate cluster) and All-In-One (AIO, putting all entities in one cluster) have almost identical results as those in the evaluation[11]. OIO can yield good performance, indicating that the names in test data are highly ambiguous. As alluded to in the title, context and categories both are very useful disambiguation features. CNTX is essentially very similar to the system presented in (Bagga and Baldwin, 1998b) and is a strong baseline[12] (outperforming $3/4$ of the systems in WePS). Note that CNTX has high purity but inferior inverse purity, indicating that using the context extracted by the IE system alone is unable to link many coreferent entities. Interestingly, we observe that using only the top-K categories (TC1+3) can also achieve competitive F score, though in a very different manner. TC1+3 recalls much more coreferent entities (significantly improving inverse purity), but at the same time also introduces noise.

Finally, adding document categories and using IE results (i.e. using all features in Table 1), our CDC system achieves 22% and 18% relative improvement compared to CNTX in F (purity) and B-CUBED scores, respectively. In particular, inverse purity improves by 46% relatively, implying that the additional evidence significantly improves the recall of coreferent entities (when there is a lack of context similarity in the traditional method). Overall, the comprehensive approach in this work outperforms the top-tiered systems in the WePS evaluation.

8 Conclusion and Future Work

This work proposes a synergy of three levels of analysis for the web cross document coreference task. On the document level, we use text categories, trained from thousands of ODP categories and over a million pages, as a concise representation of the documents. Categorization is a robust strategy for coreferencing web documents with diverse topics, formats and when there is a lack of extraction coverage or word matching. Two types of sub-document level evidence are also used in our approach. First, we apply an information extraction system to extract attributes and relations of named entities from the documents and perform within document coreference. Second, we use the context of the entities, a natural outcome of the IE system as a focused description of the named entity that may miss the extraction process. A CDC system has been implemented based on the IE and the text categorization components to provide a comprehensive solution to the web CDC task. We demonstrate the importance of each component in our system and benchmark our system with the WePS dataset which shows superior CDC performance.

There are a number of interesting directions for future research. Recently, Open IE was proposed in (Etzioni et al., 2008) for Web information extraction. This can be a more powerful alternative to traditional IE toolkits for Web CDC, though measuring the semantic similarity for a vast variety of relations can be another research issue. Employing external background knowledge such as Wikipedia (Han and Zhao, 2009) while maintaining scalability can also be an orthogonal direction for further improvement.

[11] Most person names in this set have only one underlying identity per document; thus the results are comparable despite the simplifying assumption of the WePS evaluation.

[12] We use context similarity 0.2 as the clustering threshold (which has the best performance in training data).

References

Artiles, Javier, 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)*, pages 64–69.

Bagga, Amit and Breck Baldwin. 1998a. Algorithms for scoring coreference chains. In *First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*.

Bagga, Amit and Breck Baldwin. 1998b. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th ACL and 17th COLING*, pages 79–85.

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

Breiman, Leo. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Chen, Ying and James Martin. 2007. Towards robust unsupervised personal name disambiguation. In *Proc. of EMNLP and CoNLL*, pages 190–198.

Crammer, Koby and Yoram Singer. 2003. A family of additive online algorithms for category ranking. *J. Machine Learning Research*, 3:1025–1058.

Crammer, Koby, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research (JMLR)*, 7:551–585.

Ester, M., H. Kriegel, J. Sander, and X. Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd KDD Conference*, pages 226 – 231.

Etzioni, Oren, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Communications of ACM*, 51(12):68–74.

Gooi, Chung H. and James Allan. 2004. Cross-document coreference on a large scale corpus. In *Proceedings of HLT-NAACL 2004*, pages 9–16.

Han, Xianpei and Jun Zhao. 2009. Named entity disambiguation by leveraging Wikipedia semantic knowledge. In *Proceedings of the 18th Conf. on Information and knowledge management (CIKM)*, pages 215–224.

Huang, Jian, Seyda Ertekin, and C. Lee Giles. 2006. Efficient name disambiguation for large scale databases. In *Proc. of 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 536 – 544.

Huang, Jian, Sarah M. Taylor, Jonathan L. Smith, Konstantinos A. Fotiadis, and C. Lee Giles. 2009. Profile based cross-document coreference using kernelized soft relational clustering. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 414–422.

Li, Xin, 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.

Liaw, Andy and Matthew Wiener. 2002. Classification and regression by randomforest. *R News*, 2(3).

Madani, Omid and Jian Huang. 2008. On updates that constrain the features' connections during learning. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 515–523.

Mann, Gideon S. and David Yarowsky. 2003. Unsupervised personal name disambiguation. In *Proceedings of the seventh conference on Natural language learning (CoNLL)*, pages 33–40.

Ng, Vincent and Claire Cardie. 2002. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pages 1–7.

Niu, Cheng, 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 598–605.

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

Treeratpituk, Pucktada and C. Lee Giles. 2009. Disambiguating authors in academic publications using random forests. In *Proceedings of the ACM/IEEE Joint Conference on Digital libraries (JCDL)*, pages 39–48.

Vapnik, V. 1998. *Statistical Learning Theory*. John Wiley and Sons, Inc., New York.

Wan, Xiaojun, 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.