Solving the "Who's Mark Johnson" Puzzle: Information Extraction Based Cross Document Coreference

Jian Huang[†*] **Sarah M. Taylor**[‡] **Jonathan L. Smith**[‡] **Konstantinos A. Fotiadis**[§] **C. Lee Giles**[†]
[†]Information Sciences and Technology
Pennsylvania State University, University Park, PA 16802, USA
[‡§]Advanced Technology Office, Lockheed Martin IS&GS
[‡]4350 N. Fairfax Drive, Suite 470, Arlington, VA 22203, USA
[§]230 Mall Blvd, King of Prussia, PA 19406, USA

Abstract

Cross Document Coreference (CDC) is the problem of resolving the underlying identity of entities across multiple documents and is a major step for document understanding.

We develop a framework to efficiently determine the identity of a person based on extracted information, which includes unary properties such as gender and title, as well as binary relationships with other named entities such as co-occurrence and geo-locations. At the heart of our approach is a suite of similarity functions (specialists) for matching relationships and a relational density-based clustering algorithm that delineates name clusters based on pairwise similarity. We demonstrate the effectiveness of our methods on the WePS benchmark datasets and point out future research directions.

1 Introduction

The explosive growth of web data offers users both the opportunity and the challenge to discover and integrate information from disparate sources. As alluded to in the title, a search query of the common name "Mark Johnson" refers to as many as 70 namesakes in the top 100 search results from the Yahoo! search engine, only one of whom is the Brown University professor and co-author of an ACL 2006 paper (see experiments). Cross document coreference (CDC) (Bagga and Baldwin, 1998) is a distinct technology that consolidates named entities *across* documents according to their real referents. Despite the variety of styles and content in different text, CDC can break the boundaries of documents and cluster those mentions referring to the same Mark Johnson. As unambiguous person references are key to many tasks, e.g. social network analysis, this work focuses on person named entities. The method can be later extended to organizations.

We highlight the key differences between our proposed CDC system with past person name search systems. First, we seek to transcend the simple bag of words approaches in earlier CDC work by leveraging state-of-the-art information extraction (IE) tools for disambiguation. The main advantage is that our IE based approach has access to accurate information such as a person's work titles, geo-locations, relationships and other attributes. Traditional IR approaches, on the other hand, may naively use the terms in a document which can significantly hamper accuracy. For instance, an article about Hillary Clinton may contain references to journalists, politicians who make comments about her. Even with careful word selection, such textual features may still confuse the disambiguation system about the true identity of the person. The information extraction process in our work can thus be regarded as an intelligent feature selection step for disambiguation. Second, after coreferencing, our system not only yields clusters of documents, but also structured information which is highly useful for automated document understanding and data mining.

We review related work on CDC next and describe our approach in Section 3. The methods are evaluated on benchmark datasets in Section 4. We discuss directions for future improvement in Section 5 and conclude in Section 6.

2 Related Work

There is a long tradition of work on the within document coreference (WDC) problem in NLP,

[*] Contact author: jhuang@ist.psu.edu

Proceedings of the NAACL HLT Student Research Workshop and Doctoral Consortium, pages 7–12,
Boulder, Colorado, June 2009.

which links named entities with the same referent *within* a document into a WDC chain. State-of-the-art WDC systems, e.g. (Ng and Cardie, 2001), leverage rich lexical features and use supervised and unsupervised machine learning methods.

Research on cross document coreference began more recently. (Bagga and Baldwin, 1998) proposed a CDC system to merge the WDC chains using the Vector Space Model on the summary sentences. (Gooi and Allan, 2004) simplified this approach by eliminating the WDC module without significant deterioration in performance. Clustering approaches (e.g. hierarchical agglomerative clustering (Mann and Yarowsky, 2003)) have been commonly used for CDC due to the variety of data distributions of different names. Our work goes beyond the simple co-occurrence features (Bagga and Baldwin, 1998) and the limited extracted information (e.g. biographical information in (Mann and Yarowsky, 2003) that is relatively scarce in web data) using the broad range of relational information with the support of information extraction tools. There are also other related research problems. (Li et al., 2004) solved the robust reading problem by adopting a probabilistic view on how documents are generated and how names are sprinkled into them. Our previous work (Huang et al., 2006) resolved the author name ambiguity problem based on the metadata records extracted from academic papers.

3 Methods

The overall framework of our CDC system works as follows. Given a document, the information extraction tool first extracts named entities and constructs WDC chains. It also creates linkages (relationships) between entities. The similarity between a pair of relationships in WDC chains is measured by an *awakened* similarity specialist and the similarity between two WDC chains is determined by the mixture of awakened specialists' predictions. Finally, a density-based clustering method generates clusters corresponding to real world entities. We describe these steps in detail.

3.1 Entity and Relationship Extraction

Given a document, an information extraction tool is first used to extract named entities and perform within document coreference. Hence, named entities in each document are divided into a set of WDC chains, each chain corresponding to one real world entity. In addition, state-of-the-art IE tools are capable of creating relational information between named entities. We use an IE tool AeroText[1] (Taylor, 2004) for this purpose. Besides the attribute information about the person named entity (first/middle/last names, gender, mention, etc), AeroText also extracts relationship information between named entities, such as Family, List, Employment, Ownership, Citizen-Resident-Religion-Ethnicity, etc, as specified in the Automatic Content Extraction (ACE) evaluation. The input to the CDC system is a set of WDC chains (with relationship information stored in them) and the CDC task is to merge these WDC chains[2].

3.2 Similarity Features

We design a suite of similarity functions to determine whether the relationships in a pair of WDC chains match, divided into three groups:

Text similarity. To decide whether two names in the co-occurrence or family relationship match, we use SoftTFIDF (Cohen et al., 2003), which has shown best performance among various similarity schemes tested for name matching. SoftTFIDF 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, spelling errors, etc.

Semantic similarity. Text or syntactic similarity is not always sufficient for matching relationships. For instance, although the mentions "U.S. President" and "Commander-in-chief" have no textual overlap, they are semantically highly related as they can be synonyms. We use WordNet and the information theory based JC semantic distance (Jiang and Conrath, 1997) to measure the semantic similarity between concepts in relationships such as mention, employment, ownership and so on.

[1] AeroText is a text mining application for content analysis, with main focus on information extraction including entity extraction and intrasource link analysis (see `http://en.wikipedia.org/wiki/AeroText`).

[2] We make no distinctions whether WDC chains are extracted from the same document. Indeed, the CDC system can correct the WDC errors due to lack of information for merging named entities within a document.

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 rule-based similarity functions to handle nicknames (Bill and William), acronyms (COLING for International Conference on Computational Linguistics), and geographical locations[3].

3.3 Learning a Similarity Matrix

After the similarity features between a pair of WDC chains are computed, we need to compute the pairwise distance metric for clustering. (Cohen et al., 2003) trained a binary SVM model and interpreted its confidence in predicting the negative class 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 availability of relationships in two WDC chains. Therefore, we treat each similarity function as a subordinate predicting algorithm (called specialist) and utilize the specialist learning framework (Freund et al., 1997) to combine the predictions. Here, a specialist is *awake* only when the same relationships are present in two WDC chains. Also, a specialist can refrain from making a prediction for an instance if it is not confident enough. In addition to the similarity scores, specialists have different weights, e.g. a match in a family relationship is considered more important than in a co-occurrence relationship.

The Specialist Exponentiated Gradient (SEG) (Freund et al., 1997) algorithm is adopted to learn to mix the specialists' prediction. Given a set of T training instances $\{\mathbf{x_t}\}$ ($x_{t,i}$ denotes the i-th specialist's prediction), the SEG algorithm minimizes the square loss of the outcome $\tilde{y}$ in an online manner (Algorithm 1). In each learning iteration, SEG first predict $\tilde{y}_t$ using the set of awake experts E_t with respect to instance $\mathbf{x}_t$. The true outcome y_t (1 for coreference and 0 otherwise) is then revealed and square loss L is incurred. SEG then updates the weight distribution $\mathbf{p}$ accordingly.

Algorithm 1 SEG (Freund et al., 1997)

```
Input: Initial weight distribution $\mathbf{p^1}$;
    learning rate $\eta > 0$; training set $\{\mathbf{x_t}\}$
for t=1 to T do
    Predict using:
        $$\tilde{y}_t = \frac{\sum_{i \in E_t} p_i^t x_{t,i}}{\sum_{i \in E_t} p_i^t} \quad (1)$$
    Observe true label $y_t$ and incur square loss $L(\tilde{y}_t, y_t) = (\tilde{y}_t - y_t)^2$
    Update weight distribution: for $i \in E_t$
        $$p_i^{t+1} = p_i^t e^{-2\eta x_{t,i}(\tilde{y}_t - y_t)} \frac{\sum_{j \in E_t} p_j^t}{\sum_{j \in E_t} p_j^t e^{-2\eta x_{t,i}(\tilde{y}_t - y_t)}}$$
        $p_i^{t+1} = p_i^t$, otherwise
end for
Output: Model $\mathbf{p}$
```

To sum up, the similarity between a pair of WDC chains w_i and w_j can be represented in a similarity matrix $\mathcal{R}$, with $r_{i,j}$ computed by the SEG prediction step using the learned weight distribution $\mathbf{p}$ (Equation 1). A relational clustering algorithm then clusters entities using $\mathcal{R}$, as we introduce next.

3.4 Relational Clustering

The set of WDC chains to be clustered are represented by a relational similarity matrix. Most of the work in clustering, however, is only capable of clustering numerical object data (e.g. K-means). Relational clustering algorithms, on the other hand, cluster objects based on the less direct measurement of similarity between object pairs. We choose to use a density based clustering algorithm DBSCAN (Ester et al., 1996) mainly for two reasons.

First, most clustering algorithm require the number of clusters K as an input parameter. The optimal K can apparently vary greatly for names with different frequency and thus is a sensitive parameter. Even if a cluster validity index is used to determine K, it usually requires running the underlying clustering algorithm multiple times and hence is inefficient for large scale data. DBSCAN, as a density based clustering method, only requires density parameters such as the radius of the neighborhood ϵ that are universal for different datasets. As we show in the experiment,

[3] Though a rich set of similarity features has been built for matching the relationships, they may not encompass all possible cases in real world documents. The goal of this work, however, is to focus on the algorithms instead of knowledge engineering.

density parameters are relatively insensitive for disambiguation performance.

Second, the distance metric in relational space may be non-Euclidean, rendering many clustering algorithms ineffective (e.g. single linkage clustering algorithm is known to generate chain-shaped clusters). Density-based clustering, on the other hand, can generate clusters of arbitrary shapes since only objects in dense areas are placed in a cluster.

DBSCAN induces a density-based cluster by the core objects, i.e. objects having more than a specified number of other data objects in their neighborhood of size ϵ. In each clustering step, a seed object is checked to determine whether it's a core object and if so, it induces other points of the same cluster using breadth first search (otherwise it's considered as a noise point). In interest of space, we refer readers to (Ester et al., 1996) for algorithmic details of DBSCAN and now turn our attention to evaluating the disambiguation performance of our methods.

4 Experiments

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

4.1 Evaluation Measures

We evaluate the performance of our method using the standard purity and inverse purity clustering metrics. Let a set of clusters $\mathcal{C} = \{C_1, ..., C_s\}$ denote the system's output and a set of categories $\mathcal{D} = \{D_1, ..., D_t\}$ be the gold standard. Both $\mathcal{C}$ and $\mathcal{D}$ are partitions of the WDC chains $\{w_1, ..., w_n\}$ ($n = \sum_i |C_i| = \sum_j |D_j|$). First, the precision of a cluster C_i *w.r.t.* a category D_j is defined as,

$$\text{Precision}(C_i, D_j) = \frac{|C_i \cap D_j|}{|C_i|}$$

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

$$\text{Purity}(\mathcal{C}, \mathcal{D}) = \sum_{i=1}^{s} \frac{|C_i|}{n} \max_j \text{Precision}(C_i, D_j)$$

Hence purity penalizes putting noise WDC chains in a cluster. Trivially, the maximum purity (i.e. 1) can be achieved by making one cluster per WDC chain (referred to as the one-in-one baseline).

Reversing the role of clusters and categories, $\text{Inverse\_purity}(\mathcal{C}, \mathcal{D}) \overset{def}{=} \text{Purity}(\mathcal{D}, \mathcal{C})$. Inverse Purity penalizes splitting WDC chains belonging to the same category into different clusters. The maximum inverse purity can be achieved by putting all chain in one cluster (all-in-one baseline).

Purity and inverse purity are similar to the precision and recall measures commonly used in information retrieval. There is a tradeoff relationship between the two and their harmonic mean $F_{0.5}$ is used for performance evaluation.

4.2 Datasets

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

We note that the annotation in WePS makes the simplifying assumption that each document refers to only one real world person among the namesakes in question. The CDC task in the perspective of this paper, however, is to merge the WDC chains rather than documents. Hence in our evaluation, we adopt the document label to annotate the WDC chain from the document 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 ($F_{0.5} = 0.61, 0.40$ respectively). Hence the performance reported here is comparable to the official evaluation results (Artiles et al., 2007).

4.3 Experiment Results

We computed the similarity features from the WDC chains extracted from the WePS training data and subsampled the non-coreferent pairs to generate a

Table 1: Cross document coreference performance (macro-averaged scores, I-Pur denotes inverse purity).

Test set	Method	Purity	I-Pur	$F_{0.5}$
Wikipedia	AT-CDC	0.684	0.725	0.687
ACL-06	AT-CDC	0.792	0.657	0.712
US Census	AT-CDC	0.772	0.700	0.722
Global Average	AT-CDC	0.749	0.695	0.708
	One-in-one	1.000	0.482	0.618
	All-in-one	0.279	1.000	0.389

training set of around 32k pairwise instances. We then used the SEG algorithm to learn the weight distribution model. The macro-averaged cross document coreference results on the WePS test sets are reported in Table 1. The $F_{0.5}$ score of our CDC system (AT-CDC) is 0.708, comparable to the test results of the first tier systems in the official evaluation. The two baselines are also included. Because the test set is very ambiguous (on average only two documents per real world entity), the one-in-one baseline has relatively high $F_{0.5}$ score.

The Wikipedia, ACL06 and US Census sets have on average 56, 31 and 50 entities per name respectively. We notice that as the data set becomes more ambiguous, purity decreases implying it's harder for the system to discard irrelevant documents from a cluster. The other case is true for inverse purity. In particular, we are interested in how the coreference performance changes with the number of entities per name (which can be viewed as the ambiguity level of a data set). This is shown in Figure 1. We observe that in general the harmonic mean of the purity is fairly stable across different number of entities per dataset (generally within the band between 0.6 and 0.8). This is important because the system's performance does not vary greatly with the underlying data characteristics. There is a particular name (with only one underlying referent) that appears to be an outlier in performance in Figure 1. After examining the extraction results, we notice that the extracted relationships refer to the same person's employment, coauthors and geo-locations. The generated CDC clusters correctly reflect the different aspects of the person but the system is unable to link them together due to the lack of information for merging. This motivates us to further improve performance in future work.

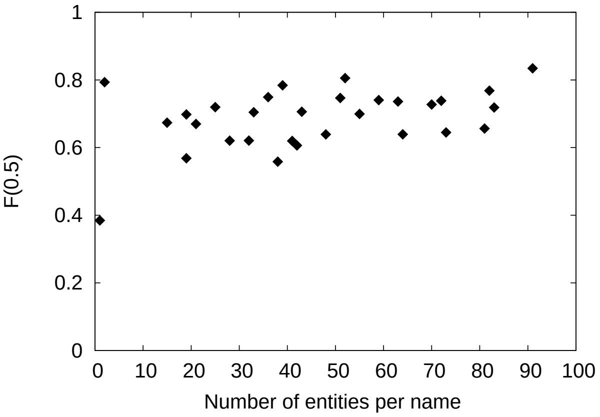

Figure 1: Coreference performance for names with different number of real world entities.

Figure 2 shows how the coreference performance changes with different density parameter ϵ. We observe that as we increase the size of the ϵ neighborhood, inverse purity increases indicating that more correct coreference decisions are made. On the other hand, purity decreases as more noise WDC chains appear in clusters. Due to this tradeoff relationship, the F score is fairly stable with a wide range of ϵ values and hence the density parameter is rather insensitive (compared to, say, the number of clusters K).

5 Future Work

We see several opportunities to improve the coreference performance of the proposed methods.

First, though the system's performance compares favorably with the WePS submissions, we observe that purity is higher than inverse purity, indicating that the system finds it more difficult to link coreferent documents than to discard noise from

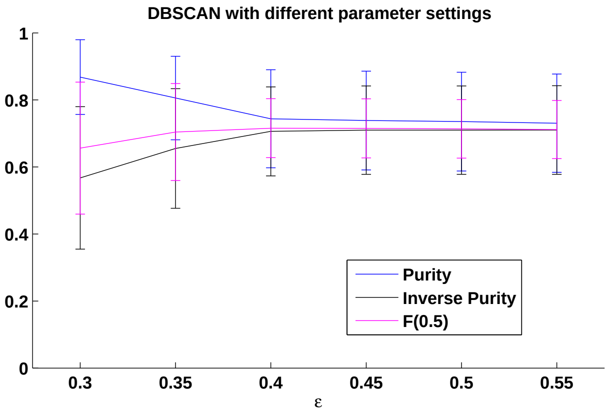

Figure 2: Coreference performance with different ϵ.

clusters. Thus coreferencing based solely on the information generated by an information extraction tool may not always be sufficient. For one, it remains a huge challenge to develop a general purpose information extraction tool capable of applying to web documents with widely different formats, styles, content, etc. Also, even if the extraction results are perfect, relationships extracted from different documents may be of different types (family memberships vs. geo-locations) and cannot be directly matched against one another. We are exploring several methods to complement the extracted relationships using other information:

• *Context-aided CDC*. The context where an named entity is extracted can be leveraged for coreference. The bag of words in the context tend to be less noisy than that from the entire document. Moreover, we can use noun phrase chunkers to extract base noun phrases from the context. These word or phrase level features can serve as a safenet when the IE tool fails.

• *Topic-based CDC*. Similar to (Li et al., 2004), document topics can be used to ameliorate the sparsity problem. For example, the topics *Sport* and *Education* are important cues for differentiating mentions of "Michael Jordan", which may refer to a basketball player, a computer science professor, etc.

Second, as noted in the top WePS run (Chen and Martin, 2007), feature development is important in achieving good coreference performance. We aim to improve the set of similarity specialists in our system by leveraging large knowledge bases.

Moreover, although the CDC system is developed in the web person search context, the methods are also applicable to other scenarios. For instance, there is tremendous interest in building structured databases from unstructured text such as enterprise documents and news articles for data mining, where CDC is a key step for "understanding" documents from disparate sources. We plan to continue our investigations along these lines.

6 Conclusions

We have presented and implemented an information extraction based Cross Document Coreference (CDC) system that employs supervised and unsupervised learning methods. We evaluated the proposed methods with experiments on a large benchmark disambiguation collection, which demonstrate that the proposed methods compare favorably with the top runs in the SemEval evaluation. We believe that by incorporating information such as context and topic, besides the extracted relationships, the performance of the CDC can be further improved. We have outlined research plans to address this and several other issues.

References

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. 2007. The SemEval-2007 WePS evaluation: Establishing a benchmark for the web people search task. In *Proc 4th Int'l Workshop on Semantic Evaluations (SemEval)*.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proc. of 36th ACL and 17th COLING*.

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

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

Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of 2nd KDD*, pages 226 – 231.

Yoav Freund, Robert E. Schapire, Yoram Singer, and Manfred K. Warmuth. 1997. Using and combining predictors that specialize. In *Proceedings of 29th ACM symposium on Theory of computing (STOC)*.

Chung H. Gooi and James Allan. 2004. Cross-document coreference on a large scale corpus. In *HLT-NAACL*.

Jian Huang, Seyda Ertekin, and C. Lee Giles. 2006. Efficient name disambiguation for large scale databases. In *Proc. of 10th PKDD*, pages 536 – 544.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of ROCLING X*.

Xin Li, Paul Morie, and Dan Roth. 2004. Robust reading: Identification and tracing of ambiguous names. In *Proceedings of HLT-NAACL*, pages 17–24.

Gideon S. Mann and David Yarowsky. 2003. Unsupervised personal name disambiguation. In *Proceedings of HLT-NAACL*, pages 33–40.

Vincent Ng and Claire Cardie. 2001. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th ACL*, pages 104–111.

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