The Workshop Programme

9:00 – 9:30 AM — Welcome and Introduction
Keith Miller, The MITRE Corporation

9:30 – 10:00 AM — *Semi-automatic Labeling of (Coreferent) Named Entities: An Experimental Study*
Erwan Moreau, Télécom ParisTech
François Yvon, Université Paris Sud & LIMSI/CNRS
Olivier Cappé, LTCI/CNRS & Télécom ParisTech

10:00 – 10:30 AM — *Adaptive Matching of Arabic Names*
Dmitry Zelenko, SRA International

10:30 – 11:00 AM — Break

11:00 – 12:15 PM — Group activity: name matching adjudication

12:15 – 12:45 PM — *Some Linguistic Considerations of Entity Resolution and Retrieval*
David Murgatroyd, Basis Technology Corp.

12:45 – 1:30 PM — Group discussion: desiderata and concepts for entity resolution evaluation

1:30 – 2:30 PM — Lunch

2:30 – 3:00 PM — *Creating a Gold Standard for Person Cross-Document Coreference Resolution in Italian News*
Luisa Bentivogli, Fondazione Bruno Kessler
Christian Girardi, Fondazione Bruno Kessler
Emanuele Pianta, Fondazione Bruno Kessler

3:00 – 4:00 PM — Group activity and discussion: adjudicating ground truth for entity resolution evaluation. How should uncertainty in the annotation be dealt with?

4:00 – 4:30 PM — Break

4:30 – 5:00 PM — *Methods for Evaluating Entity Disambiguation*
Matthias Blume, Fair Isaac Corporation
Paul Kalmar, Fair Isaac Corporation

5:00 – 5:30 PM — Group discussion: evaluation metrics

5:30 – 6:00 PM — *Linking, Mapping, and Clustering Entity Records in Information-Based Solutions for Business and Professional Customers*
Jack G. Conrad, Tonya Custis, Christopher Dozier,
Terry Heinze, Marc Light, Sriharsha Veeramachaneni
Thomson Corporation

6:00 – 7:00 PM — Group activity and discussion: task-based evaluation. Wrap up.

Workshop Organisers

Keith J. Miller (The MITRE Corporation)
Mark Arehart (The MITRE Corporation)
Sherri Condon (The MITRE Corporation)
Jason Duncan (U.S. Department of Defense)
Louise Guthrie (University of Sheffield)
Richard Lutz (The MITRE Corporation)
Massimo Poesio (Universitá di Trento)

Table of Contents

Author Index

Semi-automatic labeling of coreferent named entities: an experimental study

Erwan Moreau, François Yvon, Olivier Cappé

Télécom ParisTech & LTCI/CNRS, Univ. Paris Sud & LIMSI/CNRS, Télécom ParisTech & LTCI/CNRS
erwan.moreau@enst.fr, yvon@limsi.fr, cappe@enst.fr

Abstract

In this paper, we investigate the problem of matching coreferent named entities extracted from text collections in a robust way: our long-term goal is to build similarity methods without (or with the minimum amount of) prior knowledge. In this framework, string similarity measures are the main tool at our disposal. Here we focus on the problem of evaluating such a task, especially in finding a methodology to label the data in a semi-automatic way.

1. Introduction

In this paper, we study the problem of matching coreferent named entities in text collections, focusing primarily on orthographical variations in nominal groups (i.e. we do not handle the case of pronominal references). As described in the literature (e.g. (Christen, 2006)), textual differences between entities are due to various reasons: typographical errors, names written in different ways (with/without first name, with/without title, etc.), abbreviations, lack of precision in organization names, etc. Among them, we are particularly interested on capturing textual variations that are due to transliterations (translations between different alphabets). Identifying textual variations in entities is useful in many text mining and/or information retrieval tasks. In the former case, it will act as a useful normalization step, thus limiting the growth of the indexing vocabulary (see e.g. (Steinberger et al., 2006)). In the latter case, for instance, it allows to retrieve relevant documents even in the face of misspelling (in the query or in the document).

There are different ways to tackle the problem of NE matching: the first and certainly most reliable one consists in studying the specific features of the data, and then use any available tool to design a specialized method for the matching task. This approach will generally take advantage of language-specific (e.g. in (Freeman et al., 2006)) and domain-specific knowledge, of any external possible resources (e.g. names dictionaries, etc.), and of any information about the entities to process (especially their type: for example, there are differences between person names and organizations). In such an in-depth approach, human expertise is required in numerous ways.

The second approach is the *robust* one: we propose here to try to match any kind of NE, extracted from "real world" (potentially quite noisy) sources, without any kind of prior knowledge[1]. One looks for coreferent NE, whatever their type, source, language[2] or quality[3]. Such robust similarity methods may be useful for a lot of generic tasks, in which maximum accuracy is not the main criterion, or simply where the required resources are not available.

The orthographic similarity between strings is usually evaluated through some sort of string similarity measure. The literature on string comparison metrics is abundant, containing both general techniques and more linguistically motivated measures, see e.g. (Cohen et al., 2003) for a review. From a bird's eye view, these measures can be roughly sorted in two classes[4]:

- "Sequential character-based methods", which look for identical characters in similar positions. The most well known is certainly the Levenshtein edit distance, for which there exists a lot of variants/improvements and efficient algorithms (Navarro, 2001); the Jaro distance is also commonly used in record linkage problems (Winkler, 1999).
- "Bag-of-words methods", which are based on the number of common words between two strings, irrespective of their position. In this category fall very simple measures like the Jaccard similarity or overlap coefficient, or more elaborated ones like the Cosine similarity applied to TF-IDF weights. A related family of measures applies the same kinds of computation to "bag of (characters) n-grams" representation.

The application of these measures is relatively well documented in the database literature (see e.g. (Winkler, 1999)); however, when dealing with named entities found in text collections, it is less clear which measure(s) should be considered (see however (Freeman et al., 2006; Pouliquen et al., 2006)). Furthermore, most work on named entity matching has focused on morphological (formal) similarity. Yet, a major difference between the record linkage application and text applications is the availability of information regarding the context of occurrences of entities. We expect that this extra-information could help solve cases that are difficult for the morphological similarity measures; a similar idea has already been used for disambiguating

[1] In this kind of knowledge are included the need for hand-tuning parameters or defining language-specific heuristics.

[2] Actually we have only studied English and French (our approach is neither "multilingual", in the sense that it is not specific to multingual documents).

[3] In particular, this task clearly depends on the NE recognition step, which may introduce errors.

[4] We omit measures based on phonetic similarity such as Soundex, because they are language-specific and/or type-specific (person names), and do not fit for text collections.

homonyms (Pedersen et al., 2005; Pedersen and Kulkarni, 2007).

Our long-term goal is to build a system for automatically detecting coreferent entities using multiple string comparison measures, through machine learning techniques to select an optimal combinations of measures. This approach however presupposes the availability of hand-labeled data, stipulating which pairs of entities are positive (coreferent), and which are negative (non-coreferent). Such data is required (i) to provide an objective criterion for selecting the best combination, and (ii) to evaluate the performance of the whole system.

As a first step in that direction, we thus present and discuss in this paper a methodology for building, in a semi-automatic manner, such a hand-labeled data. This methodology assumes that the only source of information comes from the corpus: in particular, we will not use any gazetteer. We will also assume that the preliminary text processing tasks have been performed, including named entity recognition, providing us with the locations of these entities in the documents. Finally, we assume that computation time is not restricted, and that it is possible to compute all the possible pairwise comparisons. This assumption is clearly unrealistic for very large data collections and in that case, one should resort to the use of *blocking*[5] techniques. However, in the context of the small corpora we have considered, such computation is indeed feasible, and enables us to study matching results independent from the bias that this filtering step may introduce.

When building a gold standard for referent named entities, two simple minded ideas should be immediately disregarded: (i) labeling all the existing pairs is clearly beyond reach, for this would require to examine n^2 pairs of entities, where n typically ranges in the thousands; (ii) performing a random sampling in the set of pairs would also be of little help: a randomly chosen pair of entities is almost always negative. In order to recover as many positive pairs as possible, we adopted the following methodology: first, a battery of similarity measure was computed for all the pairs of entities; the top n matches for all measure were then examined and manually labeled. This allowed us to systematically compare the matches provided by each (type of) measure. This approach was successively applied on two different corpora: based on the outcome of our first experiment, we had to somewhat refine the labeling guidelines, and extend the automatic labeling tools.

This paper is organized as follows: in Section 2., we introduce the corpora, tools and guidelines that have been used to produce a golden set of matched entities. In Section 3., we provide and discuss the results of these experiments, before concluding in Section 4..

[5]In brief, blocking consists in clustering in a first step the whole set of entities, in such a way that potentially coreferent entities belong to the same cluster and that the number of entities in each cluster is minimal. This step is intended to avoid the global quadratic comparison over the whole set of pairs, needed otherwise. The question of blocking is itself very important in record matching problems (Bilenko et al., 2006).

2. Data, approach and experiments

2.1. Input data

The first corpus we used, called "Iran nuclear threat" (INT in short), is in English and was extracted from the NTI (*Nuclear Threat Initiative*) web site[6], which collects all public data related to nuclear threat. It mainly contains news, press articles and official reports obtained from various (international) sources. This corpus, limited to the 1991-2006 years, is 236,000 words long (1.6 Mio). It was chosen because

- it contains informations from various sources, a diversity that guarantees the existence of orthographic variations in named entities,
- it focuses on Iran and is thus bound to contain many transliterated names (from Persian or Arabic)

This data is slightly noisy, due to the variety of sources and/or extraction errors. We used GATE[7] as the named entities recognizer. Recognition errors are mainly truncated entities, over-tagged entities, and common nouns beginning with a capital letter. We restricted the set of entities only to those belonging to one of the three categories: locations, organizations and persons (as recognized by GATE). We obtained this way a set of 35,000 (occurrences of) entities. We finally decided to work only on the set of entities appearing at least twice, resulting in a set of 1,588 distinct entities accounting altogether for 33,147 occurrences.

Our second corpus, called "French speaking medias" (FSM in short), is a 856,000 words long corpus, extracted from a regular crawling of a set of French-speaking newspapers web sites during a short time-frame (in July 2007). The web sites were chosen based on the following criteria: geographic diversity, large volume of content, ease of access. Once again, we made sure to include a large number of web sites from North Africa, a potential source of transliterated Arabic names.

The extraction was performed by Pertimm[8]. The tagging of named entities in the corpus was then performed by Arisem[9], recognizing a total of 34,000 occurrences of entities recognized as locations, persons or organizations. Once again, the recognition step is noisy, but significantly less so than with the English corpus: less truncated or over-tagged entities, but slightly more false entities (mainly common nouns; the latter is easier to deal with than the former: for evaluation purposes, false entities have simply to be discarded). In the following, we will only work on the set of entities appearing at least twice, which yielded a unique set of 2,533 "real" entities, corresponding to 23,725 occurrences.

2.2. Methodology

Our string matching system is intended to test, evaluate, and compare as much as possible all available similarity measures. Overall, we experimented with 48 different measures, 20 of which where imported from existing

[6]`http://www.nti.org`
[7]`http://gate.ac.uk`
[8]`http://www.pertimm.com`
[9]`http://www.arisem.com`

open source packages: SimMetrics[10] by S. Chapman and SecondString[11] by W. Cohen, P. Ravikumar and S. Fienberg. Following (Christen, 2006), (Cohen et al., 2003), (Bilenko et al., 2003), we mainly considered the following measures[12]:

- *Sequential character-based:* Levenshtein, Jaro, Jaro-Winkler, Needleman-Wunch, Smith-Waterman and variants.
- *Bag of words:* Cosine, Jaccard, Overlap (simply using the number of common words between two strings), cosine with TF-IDF weighted vectors of words.
- *N-grams-characters based (for n=1,2,3):* Jaccard-type, cosine with TF-IDF weighted vectors of n-grams.
- *Combinations of measures:* Monge-Elkan, Soft-TFIDF (proposed in (Cohen et al., 2003)).
- *Context based:* this measure correspond to the Cosine of the TF-IDF vectors representing the context of two entities; context vectors contain all the occurrences of the words occurring within a fixed distance of each entity.

Given an annotated corpus, our system performs the following computations:

1. Read the NE data and the reference dataset (whenever available), select a subset of entities to process.
2. Compute the whole matrix of measures for all (selected) entities and measures[13]: each measure is applied to every pair of entities yielding $n \times (n-1)/2$ scores.
3. Manually tag top ranking pairs as positive or negative (optional).
4. For each measure, compute the k best pairs (for a predefined value of k). For several predefined values $m \leq k$, it is then possible to evaluate the individual performance of each similarity measure, using the traditional precision/recall/f-measure metrics. Additionally, it is possible to assess how each measure behaves with respect to parameters like length, number of words or frequence.
5. For every pair of measures, we finally compute the correlation coefficient and the number of common [positively labeled] pairs in the m best scores.

[10]`http://www.dcs.shef.ac.uk/~sam/simmetrics.html`

[11]`http://secondstring.sourceforge.net`

[12]A detailed description of these measures may be found on S. Champan's web page: `http://www.dcs.shef.ac.uk/~sam/simmetrics.html`.

[13]We do not distinguish entities by type (persons, locations, organizations), because type errors are rather frequent in both corpus, therefore comparing only entities having the same type would miss some positive pairs (for example, in our datasets different occurrences of the same NE are sometimes labeled with different types).

2.3. Semi-automatic labeling

As explained above, it would be very costly to manually label as match (positive) or non-match (negative) the whole set containing $n \times (n-1)/2$ pairs, for the observed values of n. A standard solution would be to label only a randomly chosen subset of pairs: in the special case of this task, this approach is ineffective, because of the disproportion between the number of positive and negative pairs. In fact our datasets only contain only respectively 0.06% (for INT) and 0.02% (for FSM) positive pairs. This is why we tried to find all the positive pairs, assuming that the remaining lot are negative. Practically, the labeling step was based only on the best pairs as identified by our set of measures. This is clearly a methodological bias (very roughly, measures are evaluated on the basis of their own predictions), but we hope to have kept the effects of this bias as low as possible. This is because the measures we used are quite diverse and do not assign good scores to the same pairs; therefore, for each measure, we expect that the potential misses (false negatives) will be matched by some other measure, thus allowing a fair evaluation of its performance. Basically this approach is close to the TREC pooling evaluation method (see e.g. (Voorhees and Harman, 1998)): the battery of measures acts as the different participating systems. Evaluation issues are further discussed in Section 3.2..

2.3.1. Labeling the INT

For the INT corpus, the labeling is based solely on the best pairs retrieved by the different measures. For each measure, our system provides the sorted set of the k best pairs, which were then proposed for human labeling in decreasing order. A minimal number of pairs is labeled for each measure (approximatly 1000), in order not to unbalance results between measures.

The guidelines we used for labeling this corpus are the following:

- *positive pairs:* two entities are considered matching if there is a "quite obvious" coreference link. Coreference is here interpreted in a rather loose sense:
  - if one of the entities is not correctly tagged (small truncation or containing too many words), they may be labeled positive provided they are clearly recognizable. Example: "*Bushehr Nuclear Plant*", "*Completing Bushehr Nuclear Plant*"
  - in some slightly ambiguous cases, two entities are considered matching if the coreference link is highly probable. For example, "*US Senate foreign relation commission*", "*Senate foreign relation commission*" is a positive pair because the corpus never talks about the "*Senate foreign relation commission*" of another country, even if such another commission may actually exist. Also, some cases of metonymy are considered positive, although this choice is certainly questionable: for instance, "Europe" and "Western Europe" are considered matching.
- *negative pairs:* two real (well formed) entities are labeled negative only if there is no doubt about their non-coreference.

- *"don't know"*[14] *pairs:* all other cases, including:
  - at least one entity is incomplete, not recognizable or ill-formed,
  - the coreference link is doubtful (potential homonymy, lack of knowledge/information from the corpus), semantic ambiguity (e.g. *"Foreign Ministry"*, *"Russian Foreign Ministry"*).

The choice of a relatively loose definition for positive pairs was guided by the concern to label a maximum amount of positive data. The manual labeling eventually yielded 805 positive pairs, 1,877 negative pairs and 3,836 "don't know" pairs.

2.3.2. Labeling the FSM

For the French corpus, labeling was more elaborated: we used the n best pairs from each measure, but also added two new methods. The first one consists in trusting transitivity relationships: if entities A and B match and entities B and C match, then entities A and C match[15]. The second one, which is more time-consuming, is a new pass over the whole set of entities. For each entity e, the n closest entities e' according to m "good" measures were also proposed for a human annotator[16]. This provides a different (complementary) viewpoint than processing the global n best pairs: this way, some pairs that could not obtain a top ranking score (this is typically the case of short entities, which are systematically over-ranked by longest ones) have a chance to be matched. The guidelines used for labeling have also been improved, based on the experience gained on the first one:

- *positive pairs:* strict coreference, at least in the corpus. The main objective is to preserve transitivity, thus it is not possible to consider "approximative coreference matching".
- *negative pairs:* strict non-coreference.
- *uncertain pairs:* this class consists is all pairs that are rejected from the positive ones but nonetheless present an important link. Some examples are: *"ONU" (UN)* and *"Conseil de sécurité" (Security Council)*, *"Russie" (Russia)* and *"Gouvernement russe" (Russian Government*).
- *eliminated entities:* all others, which consist mostly in ill-formed entitites, but also a few special ambiguous cases.

Compared with the first corpus, more time has been spent looking for possible matches in the set of entities. For example, a lot of acronyms were manually matched against their development[17] and several special cases like *"Quai d'Orsay" and "Ministère des affaires étrangères"*[18] were also addressed. Finally, the use of a supplementary processing pass allowed to label a handful of additional positive pairs (approximately a dozen among around 30,000). For all these reasons, we think that the probability for a positive pair not to be labeled is very low. We finally labeled 741 positive pairs, 32,348 negative pairs and 419 uncertain pairs. 745 entities were discarded as ill formed in the process.

3. Experiments and discussions

Performances are evaluated under the following hypotheses, in agreement with our manual labeling procedure (see above): any unlabeled pair is considered as a negative one, and any pair marked as "don't know" (or uncertain) is simply ignored.

3.1. Main observations

Overall, all measures proved to behave similarly on both corpora. Differences are nonetheless observed between the achieved performance, which are significantly worst in the case of French-speaking medias corpus. As explained above (see parts 2.3.), this is mainly due to the fact that our labeling guidelines were more strict with this second corpus.

Measures that seem to perform best are "bag of words" measures, which compute a score given the number of common (identical) words between the two strings. As expected, taking into account the IDF (Inverse Document Frequency) gives slightly better results, that is why Cosine computed over TF-IDF weighted vectors (of words) is globally the best mesure. This seems to indicate there is a pay-off in working directly with words (as opposed to characters, n-grams characters and/or positional parameters) when comparing named entities. It is indeed true that most named entities of interest, be they person or organization names, tend to correspond to morphologically complex units (title/function+first name+last name for persons, nominal groups for organizations). Yet, this result is not entirely expected, as the Cosine distance between entities is very sensitive to small orthographic differences.

In fact, it appears that in the subset of the more easily matched pairs (pairs that appear very often as one of the best scores with any measure), sequential character-based methods perform better. This subset mostly contains pairs of long strings that only differ by one or two characters. Therefore, these pairs will eventually be also matched by word-based methods, as they also contain more words than the average (they are long), and several of which are indeed identical. These pairs will thus be matched by any measure. The main problem with character-based methods is that they have a hard time sorting out the more difficult cases.

By contrast, characters n-grams measures, particularly for n=2,3, achieve an overall better level of performance. An examination the best ranking pairs for these measures reveals that they combine features from bag of words and

[14] This category is distinct from the (really) unlabeled pairs, because it does not contain any positive or negative pair.

[15] Similarly, if A and B match but A and C do not, then B and C do not match.

[16] In practice, we used $n = 3$ and $m = 4$.

[17] Although this kind of match is out of the scope of textual similarity measures, so we do not expect to catch them.

[18] *"Quai d'Orsay"* is the address where the French Ministry of Foreign Affairs is located, and is very often used as a metonym for the Ministry.

Table 1: Positive pairs by frequence

	INT	FSM
frequence ≥ 2	805	741
frequence ≥ 3	386	421
frequence ≥ 5	202	212
frequence ≥ 10	64	72

sequential character based methods: they catch minor differences more easily than bag of words measures, but have two drawbacks: firstly, as the other ones, they favour long strings (because probability to find common n-grams is higher). Secondly, they are sometimes "confused" by long strings containing similar n-grams in a different order, thus bringing a bit more false positive than bag of words measures.

Finally, the context-based measure is a very poor individual measure. As expected, good scores are obtained for entities which have an important semantic link. But this is not precise enough to match coreferent entities: typically, an organization may be matched with the person who is its main representative. A lot of other false positives are found, such as *"Israel"* and *"Palestine"*. However, the rare true positive found are interesting, because some of them could not be found by any textual measure (like acronyms and their development). This is why we plan to use the context measure in conjunction with other measures, hoping that in this case, it will prove more useful than used in isolation.

Overall, all the (good) measures tested tend to favour long strings: the average lengths in our corpora are respectively about 13 and 11 characters long (1.9 and 1.8 words long), whereas the average length among 500 best scores for all measures is respectively 15.4 and 13.1 characters long (2.1 words long for both). We also note that the average frequency of high ranking pairs is very high compared to the global average frequency. This may be due to the fact that very frequent entities are more likely to appear with variations (observing matched pairs corroborates this hypothesis).

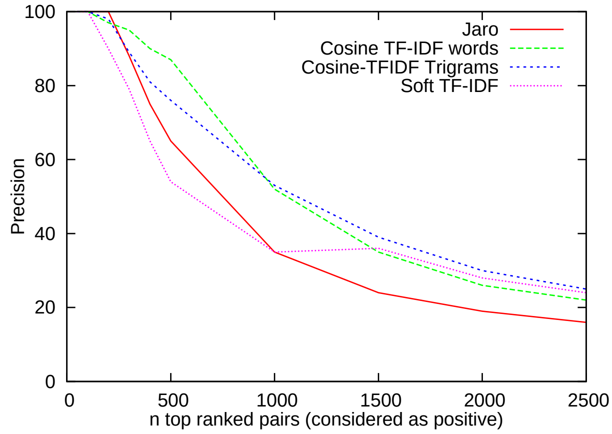
In our corpora, the most frequent sources of variation can be roughly classified as follows:

- Small typographical differences about spaces, diacritic signs, upper case letters. For example, in the FSM corpus *"Al Qaïda"* appears under 7 variations (with *i* or *ï*, with or without the hyphen, with or without uppercase *A*). These variations are easily captured by sequential character based or n-grams based methods.

- Omissions are very frequent in organization names, as in *"United States"* and *"United States of America"*, or in *"Conseil de Sécurité [de l'ONU / des Nations Unies]"* (*[UN / United Nations] Security Council*), where a PP modifier is omitted. Bag of words methods generally perform well on this kinds of pairs.

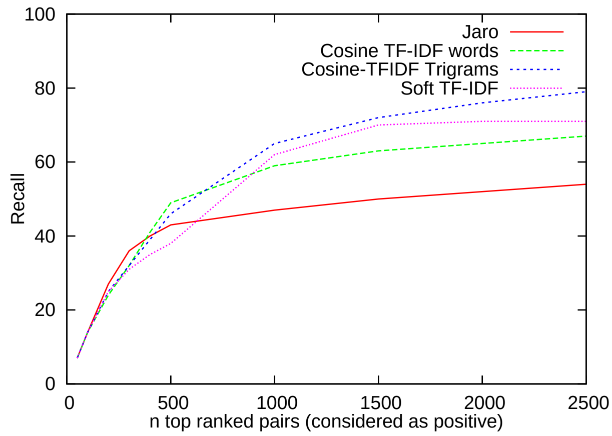
- Person names with or without the first name are also very frequent.

Figure 1: Precision (FSM)



Precision for four string comparison measures.

Figure 2: Recall (FSM)



Recall for four string comparison measures.

- Geographical orthographic variations may be more or less complex to identify, ranging from the simple pair *"Darfur"* and *"Darfour"* to the more challenging pair *"parc national d El Kala"* / *"parc naturel de la Calle"*.

Overall, all these variations are well taken care of, at least by one family of measures. More difficult cases occur when several sources of variations are combined, e.g. a change in a person name accompanied by the deletion of the first name as for the pair *"Lugovoï"* / *"Andreï Lougovoï"*.

Unsurprisingly, false positive pairs are entites that are orthographically similar but do not match, like *"ministère chinois des Affaires étrangères"* and *"Ministère russe des affaires étrangères"* (*Chinese/Russian Ministry of Foreign Affairs*) or *"South Africa"* and *"South America"*.

3.2. Discussion

The main pitfall in evaluating entities matching techniques in this framework is the disproportion between positive and negative data, together with the fact that it is (almost) im-

possible to label the whole data. As described in part 2.3., the method used to catch positive pairs depends on measures themselves. This means that there might remain some unlabeled positive pairs, which are wrongly counted as negative ones in the evaluation. This does not affect the computed precision, since enough pairs have been labeled among good scores for each measure. But recall should be interpreted with this potential bias in mind, since it depends on the number of false negative which may be underestimated.
We have tried to quantify this effect by manually searching the 2,533 unique entities in FSM for unlabeled positive pairs. As expected most of those found did not present textual similarity (otherwise they would eventually have been detected by similarity measures). Most of them were acronyms, but some other examples are also worth mentioning: *"M. Ban"* and *"Ban Ki Moon"*, *"aéroport Congonhas"* and *aéroport international de Sao Paolo"* (*Sao Paulo International Airport*), *"USA"* and *"États-Unis"* (*United States*). Under the hypothesis that we did not forget any pair, we can roughly express the probability that a positive pair remains undetected by our procedures is about 5%. A last note is in order: in all our experiments, we only considered those words that actually occurred at least twice: orthographic variations due to typos, which typically occur only once, are probably underestimated.
One of the questions we studied carefully concerns the length of entities. All (good) measures favour long strings, therefore it is possible that some pairs of short entities are missed. We have looked for best scores among short strings, in particular by filtering only entities containing only one or two words. We also studied how the distribution of the length of strings behaves with respect to the scores for several measures. Although this can not replace a systematic labeling, our observations suggest that there are simply less matching pairs with short entities, because possible textual variations are naturally proportional to the string length.
Finally, the case of uncertain pairs is also worth discussing. In our experiments, these were simply ignored; a fairer evaluation of name entity match should take them under consideration, using an intermediate status between positive and negative. For example, the pair *"ministère des Affaires étrangères"*, *"ministère français des Affaires étrangères"* (*"Ministry of foreign affairs"*, *"French Ministry of foreign affairs"*) is uncertain, although most occurrences of the general form concern the French Ministry. This question is related to another one: what is the limit for a pair to match ? Even if all occurrences of *"Ministry of foreign affairs"* in the corpus refer to the French one, should one consider this pair as a match or consider the question in a more general context: the latter viewpoint has the advantage to permit to accumulate knowledge (e.g. for large dynamic databases), contrary to the former.

4. Conclusion and future work

In this paper, we have proposed a methodology for semi-automatically labeling data in a NE matching problem, and studied the problems that arise from this methodology. We have shown that this task, which consists in finding coreferent entities extracted from corpus, presents the following peculiarities:

- very small set of positive pairs compared to the whole set of possible pairs (0.02% and 0.06% in our corpora). This problem makes it hard to obtain a sufficient amount of labeled data, thus introducing potential evaluation issues.

- some string similarity measures perform well, but no unique (existing) measure seems able to capture the variety of observed phenomena. Taking only one individual measure to compare entities requires either to make a compromise between precision and recall performance or to rely to a post-processing human validation step (as used in a lot of real systems, such as (Pouliquen et al., 2006)).

As a side note, it is worth mentioning that most sources of variations are captured by at least one family of measures.
In the future, we therefore plan to investigate methods for combining several measures, in order to improve the overall matching performances. There are different ways to do so: the first one is to use supervised learning techniques, using the now available sets of labeled data. One may also try to build new measures that would be more suited to the NE matching problem, since most existing measures are simply *string* similarity measures. In particular, it seems especially relevant to investigate unsupervised learning, or at leat semi-supervised learning techniques (for example, asking user to label only a limited number of chosen pairs).

Acknowledgements

This work has been funded by the National Project Cap Digital - Infom@gic. We thank Loïs Rigouste (Pertimm) and Nicolas Dessaigne and Aurélie Migeotte (Arisem) for providing us with the annotated French corpus.

5. References

Mikhail Bilenko, Raymond J. Mooney, William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. 2003. Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5):16–23.

Mikhail Bilenko, Beena Kamath, and Raymond J. Mooney. 2006. Adaptive blocking: Learning to scale up record linkage. In *ICDM*, pages 87–96. IEEE Computer Society.

Peter Christen. 2006. A comparison of personal name matching: Techniques and practical issues. Technical Report TR-CS-06-02, Department of Computer Science, The Australian National University, Canberra 0200 ACT, Australia, September.

William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. 2003. A comparison of string distance metrics for name-matching tasks. In Subbarao Kambhampati and Craig A. Knoblock, editors, *Proceedings of IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03), August 9-10, 2003, Acapulco, Mexico*, pages 73–78.

Andrew Freeman, Sherri L. Condon, and Christopher Ackerman. 2006. Cross linguistic name matching in english and arabic. In Robert C. Moore, Jeff A. Bilmes, Jennifer Chu-Carroll, and Mark Sanderson, editors, *HLT-NAACL*. The Association for Computational Linguistics.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88.

Ted Pedersen and Anagha Kulkarni. 2007. Unsupervised discrimination of person names in web contexts. In *Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, february.

Ted Pedersen, Amruta Purandare, and Anagha Kulkarni. 2005. Name discrimination by clustering similar contexts. In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing 2005, Mexico City, Mexico, February 13-19, 2005, Proceedings*, volume 3406 of *LNCS*, pages 226–237. Springer.

Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, Irina Temnikova, Anna Widiger, Wajdi Zaghouani, and Jan Zizka. 2006. Multilingual person name recognition and transliteration. *CORELA - Cognition, Representation, Langage.*, September 11.

Ralf Steinberger, Bruno Pouliquen, and Camelia Ignat. 2006. Navigating multilingual news collections using automatically extracted information.

Ellen M. Voorhees and Donna Harman. 1998. The text retrieval conferences (trecs). In *Proceedings of a workshop on held at Baltimore, Maryland*, pages 241–273, Morristown, NJ, USA. Association for Computational Linguistics.

W. E. Winkler. 1999. The state of record linkage and current research problems. Technical Report RR99/04, US Bureau of the Census.

Adaptive Matching of Transliterated Arabic Names

Dmitry Zelenko

SRA International
4300 Fair Lakes Ct.
Fairfax, VA 22033
dmitry_zelenko@sra.com

Abstract

We present an adaptive approach to matching transliterated names: given a training corpus of matching names, the we learn a distance function defined in terms of costs of matching ngram pairs. We evaluate name matching in the context of name retrieval and consider several evaluation metrics. We experimentally compare the new approach to the edit distance method on a large dataset of transliterated Arabic name variants.

1. Introduction

Foreign name matching is an important practical problem in information retrieval and integration: names transliterated and translated from foreign languages often exhibit a large number of orthographic variations. Therefore, integrating data sources with foreign names or searching for a foreign name requires intelligent name matching – the process that determines whether different names are likely to correspond to the same entity.

For example, the three Arabic names, *dhu-al-faqari*, *zolfaqari*, *zolfog ary* are different versions of the same name, and many Arabic names can have dozens or even hundreds possible English spellings.

We present an adaptive approach for translated and transliterated name matching. Given a training corpus of matching names, the algorithm learns a distance function defined in terms of costs of matching ngram pairs.

We present and investigate a number of evaluation metrics for name matching in the context of name retrieval. We consider several application scenarios and highlight appropriate evaluation metrics. In general, given a testing corpus of matching name variants, for each name in the corpus we use the matching model of the distance function to find the closest set of names among all other names in the corpus, with respect to the matching model. We compare the set of names to the set of true matching names, for a given name, whereby we compute recall, precision, and F-measure. We use F-measure as our evaluation metric.

For evaluation, we use a large dataset of Arabic names consisting of tens of thousands of Arabic name variants in English. We experimentally compare our approach to the edit distance method and show that the adaptive method significantly outperforms the edit distance method.

2. Preliminaries

In the following sections, we use s, x, y, z, u, v, w to denote arbitrary strings over a finite alphabet Σ, and we use a, b, c to represent letters of the alphabet. We denote the length of any string s as $|s|$ and use s_i for the ith character of s for $i \in \{1 \dots |s|\}$. We denote $s_{i \dots j} = s_i s_{i+1} \dots s_j$, which is an empty string ϵ if $i > j$. Any $s_{1 \dots i}$ is a prefix of s, and any $s_{i \dots |s|}$ is a suffix of s.

We use S to represent a set of matching strings: $S = \{s_1, \dots, s_{|S|}\}$, where $|S|$ is the cardinality of the set S. We call S a *matching set*.

A training (testing) dataset D is a collection of matching sets: $D = \{S_i : i = 1 \dots |D|\}$. We denote by S_D the set of all strings in D: $S_D = \cup S_i$.

Let $d : \Sigma^* \times \Sigma^* \to \Re$ be a distance function defined over strings. During the evaluation, for each string $s \in S_i$ we seek a set $C(s, d, \theta)$ of closest strings in S_D with respect to d, that is, $C(s, d, \theta) = \{x \in S_D : d(s, x) \leq \theta\}$ for some distance threshold θ.

Our goal is to find a distance function d and a threshold θ that for $s \in S_i$ minimize the differences between S_i and $C(s, d, \theta)$. We quantify the differences using several evaluation metrics that we present in Section 5..

We consider distance functions that are defined in terms of of operations $\delta(z, w) = t$ between pairs of strings that transform x into y with the cost of t. The cost of a sequence operations is the sum of the costs of individual operations, and after a string z is converted into w no further operations can be done on w. The distance $d(x, y)$ between x and y is defined as the minimum cost of a sequence of operations that convert x into y. In other words, $d(x, y)$ induces a minimum cost monotone alignment between x and y.

A simple example of such a distance function is the Levenstein or edit distance (Levenstein, 1965), which is defined by three types of operations: insertions ($\delta(\epsilon, a) = 1$), deletions ($\delta(a, \epsilon) = 1$), and substitutions ($\delta(a, b) = 1$).

We will consider general distance functions where edit operations $\delta(z, w)$ are defined for strings z, w of arbitrary length.

2.1. Text Indexes

We will use two types of text indexing structures: a PATRICIA tree and a generalized suffix tree.

A PATRICIA tree $PT(X)$ (Morrison, 1968) is a data structure that stores a set of strings $X = \{x_1, \dots, x_N\}$. It is a compressed representation of a trie of X. In the trie of X, each node corresponds to a distinct prefix in X. If x and xa are two prefixes in X then xa is a child of x and there is an edge (x, xa) labeled with the character a. A PATRICIA tree is a transformation of the trie of X where each node having only one child is removed, and resulted combined edges are labeled with strings that are concatenations of the corresponding characters.

A generalized suffix tree $ST(X)$ (Gusfield, 1997) for a set of strings X is a PATRICIA tree built over all suffixes of X. If we store the strings X in a separate array and for each edge of the suffix tree maintain a pointer in the array corresponding to the location of the edge label, then the resulted suffix tree representation takes $O(n)$ space and can be built in $O(n)$ time, where n is the combined length of all strings in X.

3. Top-down Learning for Name Matching

Let D be a training dataset consisting of matching sets $\{S_i\}$. Our first step is to build a generalized suffix tree $ST(S_D)$ of all strings in D. Each string $s \in S_D$ corresponds to a node in $ST(S_D)$.

Let x, y be strings in some matching set $S \in D$. We call such strings a matching pair and add them to the set M of matching pairs. Initially, the set M of matching pairs is the set of all pairs of strings present in some matching set in D.

Let $pref(x)$ be the set of nodes lying on the path from the root of the suffix tree $ST(S_D)$ to the node x excluding both the root and x. Each node $z \in pref(x)$ corresponds to a non-empty prefix of x. We denote by $suf(x)$ the set of suffixes w, such that $x = zw$ and $z \in pref(x)$, and we will write $comp(x, z)$ for w, and $comp(x, w)$ for z.

Now if x and y are a matching pair, and they have the same prefixes (suffixes), then their suffixes (prefixes) are likely to match as well. Therefore, we will add the pair of complementary non-empty suffixes or prefixes to the set M of matching pairs and apply the same process to them too. We will denote the set of common prefixes $pref(x, y) = pref(x) \cap pref(y)$ and the set of common suffixes $suf(x, y) = suf(x) \cap suf(y)$. Note that we only work with nodes of $ST(S_D)$, we never compare strings themselves – once the suffix tree is built, augmenting the set of matching pairs involves only traversing the nodes in the tree and performing integer comparisons. The algorithm TD is shown as Algorithm 1.

Algorithm 1 Algorithm TD

1: Build the suffix tree $ST(S_D)$ of all strings in D
2: Initialize the set M of matching pairs:
3: $\quad M_0 = \{\ (x, y) : \exists S_i,\ x \in S_i, y \in S_i\}$
4: $\quad M = M_0$
5: Count(x,y) = $|\{S_i,\ x \in S_i, y \in S_i\}|$
6: Count(x) = $\sum_y$Count(x,y)
7: **for all** $(x, y) \in M_0$ **do**
8: $\quad$ **while** $pref(x, y) \cap suf(x, y) \neq \emptyset$ **do**
9: $\quad\quad$ Pick $z \in pref(x, y) \cap suf(x, y)$
10: $\quad\quad$ $v = comp(x, z)$, $w = comp(y, z)$
11: $\quad\quad$ Count(z)++, Count(v)++, Count(w)++
12: $\quad\quad$ $M = M \cup \{\ (v, w)\ \}$
13: $\quad\quad$ Count(v,w)++
14: $\quad\quad$ $x = v$, $y = w$
15: $\quad$ **end while**
16: **end for**

In our implementation, we randomly pick z in step 9 from the set of matching suffixes/prefixes. This makes the time complexity of steps 8-9 linear in the depth of the suffix tree. In practice, the average depth a suffix tree is much less than the average length of strings in the suffix tree whereby we get significant speed boost.

Note that the suffix tree implementation, in addition to superior speed, also imposes a bias on the set of matching pairs restricting them to nodes in the suffix tree. This naturally eliminates many long strings from consideration.

The output of Algorithm 1 is the counts of matching pairs and counts of strings. We use the counts to compute conditional probabilities $p(x|y) = \frac{Count(x,y)}{Count(y)}$, and define the probability of match $p_m(x, y)$ to be the maximum of conditional probabilities:

$$p_m(x, y) = \begin{cases} \max\left(p(x|y), p(y|x)\right) & \text{if } x \neq y \\ 1 & \text{if } x = y \end{cases}$$

Finally, we define the matching cost $\delta(x, y)$ of the distance function to be the negative logarithm of the probability of match:

$$\delta(x, y) = -\log p_m(x, y)$$

4. Approximate Search

For evaluation, we develop an approximate search component that given a string s, a distance function d, a threshold θ, and an index of S_D (the set of all strings in a testing set D) returns a set $C(s, d, \theta)$ of closest strings in S_D with respect to D.

We use a PATRICIA tree $PT(S_D)$ to represent the index of S_D. We incrementally construct $C(s, d, \theta)$ by conducting a beam search in S_D. We architect the beam search by maintaining multiple beams, one beam for each prefix of s. Each beam maintains a set of search states, where a search state corresponds to a position within the PATRICIA tree $PT(S_D)$. A position with $PT(S_D)$ is a prefix of some string in S_D, it can be either a node of $PT(S_D)$ or lie on some edge of $PT(S_D)$.

During the beam search, we proceed incrementally: we examine the beam corresponding to a prefix s of length i (initially, $i = 0$ and its beam has only one state corresponding to an empty string). We select a state from the beam and extend it in $PT(S_D)$ using the distance function d. We generate new states for the extensions and add them to corresponding beams. At the end of the search, when $i = |s|$, the final beam will contain an approximation of $C(s, d, \theta)$.

5. Experimental Evaluation

We evaluate accuracy of name matching using the following search scenario. For each name $s \in S_i$ in a testing set, for we use a distance function to find the closest set of names among all other names in S_D, with respect to the distance function. We propose several evaluation metrics in this setting.

Our first evaluation metric does not address the problem of determining the correct threshold θ: for each $s \in S_i$ we find the set $C(s, d)$ of $|S_i|$ closest set of names in the testing set.[1] We then compute recall, precision, and F-measure of

[1]Note that the cardinality of $C(s, d)$ may be less than $|S_i|$, because the beam search produces only an approximation of $C(s, d)$. Also, some strings may not be reachable from s using the matching model of the distance function d.

Beam	ED	TD
5	32.4	70.3
10	37.3	73.0
20	41.7	75.0
50	50.4	76.1
100	52.7	76.1
200	55.1	76.0
1000	56.4	76.0

Table 1: Arabic Name Matching Performance (Fm).

$C(s, d)$ with respect to S_i:

$$R = \frac{\sum_{S_i} \sum_{s \in S_i} (|C(s,d) \cap S_i| - 1)}{\sum_{S_i} \sum_{s \in S_i} (|S_i| - 1)}$$

$$P = \frac{\sum_{S_i} \sum_{s \in S_i} (|C(s,d) \cap S_i| - 1)}{\sum_{S_i} \sum_{s \in S_i} (|C_1(s,d)| - 1)}$$

$$Fm = \frac{2PR}{P + R}$$

We substract 1 in the above formulas to exclude the search string s itself from $C(s, d)$. We use Fm as the evaluation metric in our experiments.

In the second evaluation scenario, we vary the distance threshold θ, find the set $C_2(s, d, \theta)$ of strings in S_D lying within the ball of radius θ with respect to s, and compute recall R_θ, precision P_θ, and F-measure Fm_θ in this setting.

In the third evaluation scenario, we modify computation of the recall metric. In particular, in some application of record retrieval by name $s \in S_i$, it is not necessary to retrieve all correct matches in a matching set S_i – at least one match is sufficient to retrieve the relevant record. Therefore, we modify the recall metric to reflect this scenario:

$$R^1_\theta = \frac{\sum_{S_i} \sum_{s \in S_i} sgn(|C(s,d,\theta) \cap S_i| - 1)}{M}$$

where M is the number of matching sets S_i, and $sgn(x) = 1$, if $x > 0$, and 0, otherwise. The formulas for the precision P^1_θ and F-measure Fm^1_θ stay the same as above.

5.1. Evaluation Data

We use a dataset of Arabic name variants for evaluation. The dataset consists of 8241 matching sets, contains 23352 name entries (that is, on average 3 names per a matching set), and 23050 unique names. We randomly split the dataset in training (60%) and testing (40%) sets.

For example, here is a typical matching set in the data: *'abdulmisih, 'abdilmissih , 'abd al masih , 'bdulmasih , 'abd-al-massih , 'abd-al-massiah*.

In our experiments, we treat names (including multi-word names) as strings of characters: no preprocessing or segmentation is performed. We note that the data contain very few examples of segment reordering, and our approach does not address the reordering issue.

5.2. Experimental Results

The experimental results according to the the first evaluation metric (Fm) are shown in Figure 1 and Table 1.

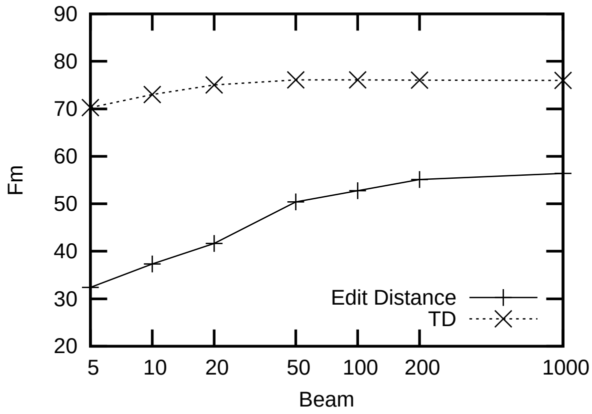


Figure 1: Arabic Name Matching Performance (Fm).

d	0.2	0.25	0.3	0.35
ED	39.1	42.4	42.4	39.9
d	1	1.1	1.2	1.3
TD	68.9	69.6	70.0	69.9

Table 2: Arabic Name Matching Performance (Fm_θ).

We compute the second evaluation metric by varying threshold and θ and computing Fm_θ for different values of the threshold. We found out experimentally that using the distance d normalized by the length of s works better than using the unnormalized distance d. Table 2 shows the performance of edit distance and top-down approaches for different values of normalized distance for the beam of 20, and makes it clear that distance functions are calibrated differently for different approaches.

Finally, we use the our third evaluation metric to compute the lookup performance statistics shown in Figure 2 and Table 3. In computing the metric, we used optimal distance thresholds (with respect to Fm^1_θ) for both the top-down distance model ($\theta = 0.8$) and the edit distance model ($\theta = 0.25$).

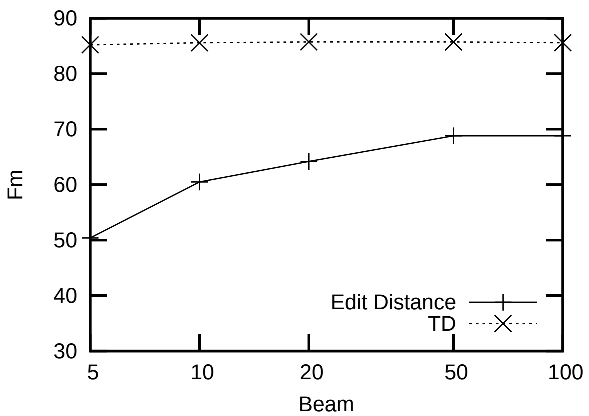


Figure 2: Arabic Name Matching Performance (Fm^1_θ).

We see from the results that the top-down algorithm significantly outperforms the edit distance approach using all of the evaluation metrics. One appealing feature is that top-

Beam	ED	TD
5	50.4	85.2
10	60.5	85.6
20	64.2	85.7
50	68.8	85.7
100	68.8	85.6

Table 3: Arabic Name Matching Performance (Fm^1_θ).

down algorithm achieves excellent performance even for small values for the beam, which makes its approximate search faster than the search using edit distance. In particular, with the performance-competitive beam of 5, the approximate search throughput with the top-down model is 4200 queries per second, while the competitive throughput for edit distance (beam=50) is only 1200 queries per second. Therefore, the top-down approach delivers not only superior accuracy but also faster search.

6. Related Work

There has been very little work on the *adaptive* name matching problem in our setting. The practical problem of culture-specific name searching and matching has been mainly addressed in the industry, where a couple of products are marketed by Basis Technologies (Basis, 2008) and IBM (IBM, 2008). Both products are carefully engineered rule-based systems, and linguistic expertise is required for their maintenance.

From the stringology perspective, there has been a lot of work on approximate string matching and searching algorithms (see, for example, (Navarro, 2001; Navarro et al., 2001) and references therein). The work mostly addresses using edit distance in search for approximate names, while we focus on generalized trainable distance metrics.

In the database community, there has been a lot of work on record linkage (see (Cohen et al., 2003) for a survey of distance metrics). We note the adaptive work on merging names and database records (Bilenko and Mooney, 2003; Bilenko et al., 2003) that aims to learn probabilistic edit distance with affine gaps for name matching. However, the edit distance is defined in terms of single characters, which makes it unlikely to work well in general cross-cultural name matching. Like our bottom-up approach, it follows (Ristad and Yianilos, 1998) in probabilistic interpretation and learning of edit distance.

In conducting approximate search we do not utilize a filtering step that uses a low cost metric to find a subset of S_D, for which the real similarity metric (e.g., edit distance) is applied. There has been a number of filtering approaches in the literature, with application to name searching, including phonetic indexing (Taft, 1970; Knuth, 1973; Gadd, 1990; Christen, 2006), ngram filtering (Cohen et al., 2003), pivot-based filtering (Chávez et al., 2001), and partition filtering (Wu and Manber, 1991). Our beam search implementation obviates the need for a preliminary filtering step and achieves good query throughput.

7. Discussion

We present a learning algorithm for name matching that is able to exploit structure inherent in a suffix tree representation of data to achieve superior accuracy. The new learning algorithm is extremely efficient (for our dataset, suffix tree construction and training take less than 1 second) and has no tunable parameters. The suffix tree representation allows to exploit matching strings of any length while enjoying the linear time computational complexity during training.

The algorithm superior performance is surprising because it discards a lot of information: pairs of matching strings in training data are ignored if they have different suffixes and prefixes. However, if the training data is plentiful then most useful matches can likely be gathered by simply splitting identical suffixes and prefixes. We experimented with an iterative version of the algorithm that uses the learned matching strings to iteratively gather additional matches. For our datasets, the iterative algorithm only slightly improves matching accuracy.

8. References

Basis. 2008. Rosette Name Indexer. http://www.basistech.com/name-indexer/.

M. Bilenko and R. Mooney. 2003. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of KDD 2003*.

M. Bilenko, W. Cohen, S. Fienberg, and R. Mooney. 2003. Adaptive name matching in information integration. *IEEE Intelligent Systems Special Issue on Information Integration on the Web*.

E. Chávez, G. Navarro, R. Baeza-Yates, and J.L. Marroquin. 2001. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, September.

P. Christen. 2006. A comparison of personal name matching: techniques and practical issues. In *Proceedings of ICDM 2006 workshop on mining complex data*.

W. Cohen, P. Ravikumar, and S. Fienberg. 2003. A comparison of string metrics for matching names and records. In *Proceedings of KDD 2003*.

T. Gadd. 1990. Phonix: The algorithm. *Program: automated library and information systems*, 24(4).

D. Gusfield. 1997. *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press.

IBM. 2008. Global Name Recognition. http://www.ibm.com/software/data/ips/products/masterdata/globalname/.

D. Knuth. 1973. *The Art of Computer Programming, volume 3: Sorting and Searching*. Addison-Wesley.

V. Levenstein. 1965. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1:8–17.

D. Morrison. 1968. PATRICIA-practical algorithm to retrieve information coded in alphanumeric. *Journal of the ACM*, 15.

G. Navarro, R. Baeza-Yates, E. Sutinen, and J. Tarhio. 2001. Indexing methods for approximate string matching. *IEEE Data Engineering Bulletin*, 24(4):19–27.

G. Navarro. 2001. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88.

E. Ristad and P. Yianilos. 1998. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532.

R. Taft. 1970. Name search techniques: New york state identification and intelligence system. Technical report, State of New York.

S. Wu and U. Manber. 1991. Fast text searching with errors. Technical Report TR-91-11, University of Arizona.

Some Linguistic Considerations of Entity Resolution and Retrieval

David Murgatroyd
Basis Technology Corp.
One Alewife Center; Cambridge, MA 02140 USA
dmurga@basistech.com

Abstract

Entity resolution and retrieval systems confront significant challenges in dealing with linguistic data such as personal names. In this paper we survey those challenges from the perspective of the data's variation, composition, distribution, under-specification, and multilinguality. We explore guidelines for integrating systems that address these challenges. We also consider strategies for evaluating such systems, including developing corpora which reflect the challenges and adopting metrics which measure how well they are met.

1. Introduction: The Tasks

Entity resolution is the task of determining if two or more given references refer to the same entity. It can be formalized as a function which maps two members of a set of references R into the set $B = \{0, 1\}$

$$resolve?: R \times R \rightarrow B$$

We may wish to view this task probabilistically, formulating it as

$$P(resolve?(r,r')=1)$$

If we posit the existence of an entity set E (e.g., the real world) where each reference in R refers to a member of E, we can recast this probability of as

$$= P(e = e' \mid r \,\&\, r')$$

where e and e' are entities referred to by r and r', respectively. This is the probability that the entities referred to by r and r' are identical given r and r'. This probability function is commutative, that is

$$P(resolve?(r,r')) = P(resolve?(r',r))$$

It is also is reflexive if the references share identity

$$P(resolve?(r,r)) = 1$$

but not necessarily if they merely share equality

$$P(resolve?(r,r')|r=r') <= 1$$

because equal records only suggest the existence of a possible common entity, not the uniqueness of a specific entity.

In practice, the entity resolution task may not be limited to identifying references to the same entity, but may also include merging those references into a single subsequent reference. A merge function may be formalized as

$$merge: R \times R \rightarrow R$$

Merging presents a number of challenges such as determining the form of the resulting reference and accommodating a merged reference in the probability model. However, we consider merging to have a large system-dependent component and do not explore it further below.

Entity retrieval is the task of determining a set of relevant references in response to a given query. Relevance for our purposes is a relation between members of a set of queries Q and set of entity references R

$$relevant: Q \times R$$

The retrieval task is: given some member of the query set, return all members of the reference set for which the relation *relevant* holds. A reference-specific perspective on this is a function which maps a single query and a single reference into the set $B = \{0, 1\}$

$$relevant?: Q \times R \rightarrow B$$

We may wish to view this task probabilistically, formulating it as

$$P(relevant?(q,r)=1)$$

We will not further decompose the concept of relevant because it is subjective, as is recognized in work addressing the broader task of information retrieval (Voorhees, 2001). Nonetheless, it is worth observing that this probability is not necessarily commutative. For example

P(relevant?("Bush", "George Bush")=1)

does not necessarily equal

P(relevant?("George Bush", "Bush")=1)

because the role of query and reference are different: the query is an expression of the user's real or perceived information need while the referent refers to a particular entity. This holds even if the query set is identical to the reference set because the fact that a reference is given as a query changes the information it conveys. We do expect this probability function to be reflexive under identity and equality. That is, a reference is always relevant for a query to which it is identical or equal.

2. Linguistic Challenges

The references and queries processed in the tasks of entity resolution and retrieval often contain linguistic data, which we define to be data arising from human natural language. While almost any string datum referring to any type of entity could fit this definition, we will focus on a specific subclass of linguistic data: personal names.

Personal names may be considered to bear much of the information useful for these tasks when applied to persons. By "personal name" we mean a label consisting of sounds and/or concepts (such as a title or honor) used to refer to a human being. A personal name may have a textual representation in a written script. We define the language-of-use of a textual representation of a name to be the language of the representation's intended audience. The name itself does not have a language-of-use but may have an identifiable language-of-origin, the language in which the name first appeared, perhaps as an adoption of a non-name word for use as a name or as an artifact of a culture which spoke that language. Identifying a name's language-of-origin is an etymological task.

As an example, "David" is a textual representation in Latin script with English as a possible language-of-use (additional possibilities include other languages natively written in Latin script) of a name whose language-of-origin is Hebrew.

The processing of personal names in these tasks encounters a number of inter-related challenges. In exploring these challenges, we will give a few examples from the representation of names in the Arabic language. Nonetheless, we believe these challenges apply to other languages as well as to other types of linguistic data beyond personal names.

2.1 Variation

A specific entity may be referred to in a variety of ways in a specific language-of-use (we will explore the challenge of multi-linguality in section 2.5). These variations may be intended or unintended by the referrer.

2.1.1. Intentional Variation

A single entity may be referred to with many names. The inventory of names may vary based on factors such as formality (e.g., nicknames) or transparency (e.g., aliases). Life events may modify the set of names for a particular entity. These events may have to do with vocation (e.g., titles) marital status (e.g., marriage/divorce/widowhood), parenthood (e.g., having a son/daughter), faith (e.g., christening, completing a pilgrimage), or life itself (e.g., posthumous names).

Even when a single specific name is considered, its representation in a specific language may evidence intentional variation for dialectical or stylistic reasons.

Some languages have multiple dialects. Arabic, for example, has both a high-prestige dialect, Modern Standard Arabic (MSA), and many low-prestige dialects. These different dialects have different phonologies which lead to differences in spelling. For example, the name which is commonly represented in MSA as قاسم has dialectical variations including آسم, كاسم, گاسم, and غاسم.

The representation of a name in a particular text may be influenced by the style of the text. For instance, some MSA texts display a clear preference for one or the other of the *hamza above* [1] and *hamza below* combining characters for typographical reasons.

[1] We refer to Unicode characters by their name in italics.

2.1.2. Unintentional Variation

Name representations processed by computational systems appear in encodings that correlate sequences of bits to characters with linguistic significance. An encoding's expressivity may enable visual ambiguities that lead to unintentional variations.

The Unicode encoding of the Arabic script contains characters with the same glyph. For instance, *farsi yeh*, for use in Persian and Urdu, has the same rendering at the beginning and in the middle of a word as the *yeh*, for use in Arabic -- both are rendered as يه when followed by *heh* and as هيه when preceded and followed by *heh*. *Farsi yeh* also has the same rendering at the end of a word as *alef maksura* -- both are rendered as هى when preceded by *heh*. Such ambiguous glyphs enable a human creating a reference via a visual feedback system to use specific a character unintentionally, especially if the input method used is unfamiliar. That is, even if a referrer is aware of the difference between two characters such as *yeh* and *farsi yeh*, he or she may not be able to determine which has been typed.

Similarly, differences in glyphs may be so minor that a human is not conscious of the difference. For instance, in Arabic script the presence or absence of dots is the sole means of visually distinguishing *teh marbuta* (ة) from *hah* (ه) and *yeh* (ي) from *alef maksura* (ى).

Also, encodings may enable composed or combined forms of characters whose visual distinction from standalone characters may be unnoticed. For example, شاطئ ends with the single character *yeh with hamza above* (although the rendering is actually that of *alef maksura*). A variation is شاطىء whose final two characters are *alef maksura* followed by *hamza*.

Along with these unintended variations stemming from encodings, unintentional variation may arise from typographical errors. That is, the character sequence entered by the referrer may not be that intended, despite potential visual feedback he or she may have received.

Different data sources may display different types and different levels of variations. Systems may want to apply different models in analyzing a reference based on the confidence in the fidelity of the data source.

2.2 Composition

A textual representation of a personal name may contain many words which may be ordered or omitted.

These individual words have a variety of sources. In addition to the life events mentioned previously, names given at birth may be given by parents, apply to the family, communicate lineage (e.g., a patronym or matronym), or derive from a geographical location (a toponym). The ordering as well as relative importance of these name words may vary by individual, data source, or culture. For example, a data source may represent a name in a "family name, given name" pattern. Similarly, in some cultures the family name often begins the name, while in others the given name often comes first. Further, some if not most of the name words may be commonly omitted in a given reference.

A data source may attempt to capture the delineation between different name components by separately representing components such as given names and patronyms. Such a decomposition is information-bearing but of limited use because of the difficulty of establishing a common delineation across data sources.

2.3 Under-specification

The textual representation of a name may contain only the information the referrer considers necessary for it to be unambiguously interpreted by a reader. Examples of this in MSA include the absence of diacritics in popular texts, the absence of white space segmentation for compound names or names ending in non-joining characters, and the use of initials or other abbreviations.

2.4 Distribution

Personal name words are an open class to which any person may add. For this reason there can be no entirely comprehensive onomasticon of name words.

Nevertheless, some name words are very common. As with many natural language phenomena, the distribution of name words appears highly non-uniform. More specifically, we analyzed a large corpus of names used in Arabic and found that the frequency of any name word is inversely proportional to its rank in a frequency table, composing a Zipfian distribution.

These two observations imply that systems may benefit from inventories of common name words but must also be prepared to handle unseen name words.

2.5 Multi-linguality

The previous challenges have focused on representations of names within a specific language. However a name may be represented in different languages-of-use. The differences in name representation may be viewed from a variety of linguistic levels.

Orthographic considerations apply to name words whose content is phonetic. This can be seen both across languages written in the same script and those written in different scripts. For instance, the name represented in MSA as أومادي may be represented in Berber as ؤمادي and in English as Umadi, reflecting the orthographies of each language and their respective relationship to phonology. Often a name's orthographic representation in one language is derived from its orthographic representation in another language via a process called transliteration. Because of possible skew in the phonetic and orthographic inventories of each language, a single name may have a variety of transliterations.

The source language for transliteration is often the name's language-of-origin. Occasionally the source language may be some more accessible language than the language-of-origin, such as if a name of Chinese origin appearing in Arabic was effectively first transliterated from Chinese to English and then from English to Chinese.

A name's etymology may affect its representation in another language-of-use. For instance the etymology given above for name represented in English as "David" also applies to the name represented as داؤود in Arabic. The English name may be represented in Arabic by that etymologically related name or by a letter-for-letter orthographic transliteration such as ديفيد.

As mentioned in the discussion of composition, the syntax of name words may vary based on the cultures or languages involved. For instance, the representation of a personal name in Arabic of Chinese origin may or may not display the family-name-first convention which its native representation in Chinese displays.

Name words whose content is conceptual such as titles (e.g., مهيب which means "field marshal") or qualifiers (e.g., الاصغر which means "the younger") are semantically translated, not transliterated. That is, they are represented using a word of the language-of-use whose meaning is closest to that intended (e.g., "Jr." for الاصغر). It is important for a system to identify which name words are likely to have been transliterated versus translated.

At the pragmatic level, references to an entity in different languages may use different names altogether based on the communication desired with the audience. For instance, an individual who may be referred to in Arabic as الأمير ("the prince") may be referred to in English using a patronymic as if it were a family name (e.g., "Mr. Laden") in an attempt to fit the naming expectations of an English-speaking audience.

3. Implementation and Integration

Systems focusing on linguistic considerations of entity resolution and retrieval may be integrated inside larger systems which also address non-linguistic considerations of these tasks. The properties of the integration affect the performance of the resulting system, as measured both in terms of the correctness of the respective tasks and in terms of the time and space resources required.

3.1 Inputs

What data should the larger system input to a sub-system which addresses linguistic aspects of these two tasks? The primary decision is between a pair-wise or set-based function. A pair-wise function considers one pair of references for resolution or one query-reference pair for retrieval. A set-based function is given access to all the references as well as the query in the case of the retrieval task (we consider batch querying outside the scope of the entity retrieval task).

Although some entity resolution algorithms such as (Menestrina et al., 2006) treat sub-components as black-box pair-wise functions, we believe a set-based function provides the best task and space/time performance. It can provide the best task performance because it has access to the relevant context across all references. It can provide the best space/time performance because it can use techniques like dynamic programming to compute partial results based on common linguistic content. It is also enables more efficient incorporation of updates to the reference set. Further, a set-based function may decrease the need for "blocking" heuristics intended to keep the number of pair-wise comparisons computationally tractable (Fellegi & Sunter, 1969).

3.2 Outputs

What output should the linguistic sub-system return to the larger system? This depends on the model the larger system uses for combining evidence from its various components. One choice is a feature-based model where the features are combined based on some machine-learned or hand-tuned function. Another choice is a probability model which treats the value returned as a true probability to be incorporated into a broader probability calculation.

Regardless of the choice of integrating model, it is important to note that the use of a similarity measurement in place of a probability model is suboptimal. For entity resolution, this is because similarity is reflexive in equality, not just in identity, in contrast to the description of resolution given above which is only reflexive in identity. As an example, consider two instances of the most common name token in Arabic. They are fully similar:

similarity("محمد", "محمد") = 1.0

But they are not fully likely to refer to the same entity:

P(resolve?("محمد ", "محمد ") << 1.0

This is because similarity measurements do not take into account the prior probabilities of the references or of the proposition that the entities are identical.

Similarities may be of some use for entity resolution, however. The probabilistic expression of the task may be decomposed into

$$= P(r \,\&\, r' \mid e=e')P(e = e') / P(r \,\&\, r')$$

$$= P(r \mid r' \,\&\, e=e')P(r' \mid e=e')P(e = e') / P(r \,\&\, r')$$

and it may be then possible to use similarity measurements to approximate the conditional probabilities, as suggested by (Blok et al, 2003.).

For entity retrieval, similarity's property of being reflexive in equality is acceptable but its commutativity is not. For example

similarity("Bush", "George Bush")=
similarity("George Bush", "Bush")

but

P(relevant("Bush", "George Bush")=1) !=
P(relevant("George Bush", "Bush")=1)

3.3 Properties

There may be properties or contracts of subcomponents of entity resolution systems which are desirable for efficient or effective integration.

For entity resolution, (Menestrina et al., 2006) show that a pair-wise function which is commutative and reflexive can be efficiently used in an entity resolution algorithm. They detail other properties for merging and data confidence.

Little has been written about desirable properties for subcomponents of entity retrieval systems. Considering the broader task of information retrieval, desirable subcomponent properties vary based on the properties of the underlying model and its mathematical basis. Some desirable properties may include ability to participate in a fast index and the ability to be represented in a vector space.

4. Evaluation

A system's performance may be measured both by the task definition as well as the computational resources required. Although evaluation of resources required is better defined and perhaps less important at this stage of the field's development, previous evaluation of the entity resolution task has focused on it (e.g., number of pair-wise comparisons made). This stems in part both from the lack of integrated systems to address the task and from the lack of evaluation corpora. The discussion below focuses on evaluation of the task itself.

4.1 Focus

The tasks of entity resolution and retrieval may process linguistic and non-linguistic data. Because of the diversity of the data types processed in each situational instance of a task, it would be premature for evaluation to focus primarily at the level of integrated systems. Rather, evaluation should distinguish, if not focus on, specific types of data (e.g., names of people). Such an emphasis would support improvement of processing needed for these different types of data.

4.2 Metrics

Entity resolution is similar to the task of coreference resolution. The former resolves members of a set of references which perhaps appear in isolation, the latter resolves members of a set of references appearing in a set of documents. The B-CUBED algorithm (Bagga & Baldwin, 1998) is widely used in coreference resolution to compute precision and recall. The ACE 2008 evaluation uses it alongside the customized ACE Value score. It improves on the prior MUC-6 metric presented in (Villain et al., 1995) because it is not sensitive to the sparseness of the resolution graph, gives credit for identifying singletons, and allows weighting at the entity or reference level. However, both B-CUBED and the MUC-6 metric rely on intersecting the reference and system resolution graphs which allows for an entity to be considered multiple times. The Constrained Entity-Alignment F-Measure (CEAF) of (Luo, 2005) avoids this by computing the optimal one-to-one alignment of the resolution graphs.

Entity retrieval is similar to the broader task of information retrieval. The former retrieves entity references based on an expression of information need about entities, the latter retrieves documents based on an expression of a more general information need. We believe the traditional measures of precision and recall used by information retrieval can be directly used by entity retrieval, including more specific versions such as precision/recall at a particular threshold or of the top-N results.

4.3 Corpora

Corpora used for entity resolution evaluation must both be representative of the linguistic challenges explored above and be annotated for ground truth. Obtaining ground truth is difficult because it is defined relative to some set of entities whose members may not be known. An obvious candidate for such an entity set is entities in the physical world, such as persons. However, few public corpora we know of are explicitly bound to unique entities in the world. Adding such bindings after the fact is difficult as it requires that the annotator be able to identify the individual in question. Non-public corpora which have such bindings may be difficult to share because of privacy or proprietary concerns.

One public source which could be fodder for an entity resolution corpus is Wikipedia. Its pages refer to real world individuals and it encodes variation in redirect links and multilinguality in other language links. However, the constant editing of Wikipedia may limit the appearance of unintentional variations such as typographical errors. Also, the specific entity type referenced by a Wikipedia page would need to be annotated. Other public databases such as citation indexes (e.g., CiteSeer) or civic records are potential sources though they may similarly contain limited variation and are largely unresolved to ground truth.

Entity resolution corpora may be adapted from corpora for other linguistic tasks. Coreference corpora are annotated for ground truth but their mention sets are often constrained to a single document in a single language. The ACE 2008 EDR corpus will address the first constraint by providing mentions from multiple documents. The ACE/ET 2007 corpus addresses the second constraint by providing multilingual mentions from translations of individual documents. No corpus addresses both of these constraints. Further, the coreferent mentions found in documents do not evidence all the challenges of entity references from a variety of sources.

Corpora can be constructed by humans generating references to an entity presented to them. If the entity is presented by speaking a name, the resulting references may display multilinguality and some variation. We have constructed small corpora via this technique, which we call a "parrot session". If the entity is presented by non-linguistic means (e.g., showing a picture) then more varied references are possible, but this requires that the audience already know one or more names for the entity.

Entity resolution corpora may be synthesized. This requires determining a number of references to generate for each entity and generating each individual reference. This may be done by sampling a generative model of entity resolution, but the resulting corpus is of course useless to evaluate the source model. If references have been mapped into a geometrical feature space, such as might be used by a discriminative model, geometric cluster generation like that explored by (Delling et al., 2006) could potentially be used. A commercial company, (Spock, 2008), sponsored an "entity resolution" evaluation where each reference was an entire document (though the property that each reference referred to exactly one entity still held). A corpus of 100,000 references was created for this evaluation, a quarter of which were annotated with ground truth. Some evidence indicates that portions of this corpus were synthesized by replacing a name in a document with a pseudoname referring to another entity to introduce ambiguity, in a method similar to that used by (Mann & Yarowsky, 2003).

Production of corpora for entity retrieval evaluation may adapt the Cranfield paradigm for information retrieval corpora creation (Cleverdon, 1997). The assumptions of Cranfield seem more tenuous when applied to entity retrieval. The assumption that relevance can be approximated by similarity is undermined by the observation that symmetry is commutative but relevance is not. The implications that all relevant entities are equally and independently so also may not hold. The additional assumptions of representative relevance across a user population and the completeness of the identified relevance sets are similarly questionable. Nonetheless adaption of a Cranfield-like paradigm may be useful for comparative evaluation as (Voorhees, 2001) found for information retrieval.

5. Conclusion

Entity resolution and retrieval are related tasks with which face similar challenges in processing linguistic data stemming from its variation, composition, distribution, under-specification, and multilinguality. The implementation of sub-systems to address these linguistic challenges should satisfy integration requirements of larger systems for entity resolution. Evaluation should distinguish, if not focus upon, these sub-systems to facilitate research. The evaluation metrics for each task should be informed by those for the related tasks of coreference resolution and information retrieval, respectively. Evaluation corpora may be difficult to produce, particularly for entity resolution, though some production methods exist.

6. References

Bagga, A., Baldwin., B. (1998). Algorithms for scoring coreference chains. In *Proceedings of the First International Conference on Language Resources and Evaluation Workshop on Linguistic Coreference*.

Blok, S., Medin, D. L., Osherson, D. (2003). Probability from similarity. Paper presented at the American Association for Artificial Intelligence Spring 2003 Symposium, Palo Alto, CA.

Cleverdon, C. (1997). The Cranfield tests on index language devices. In K. S. Jones and P. Willett, editors, *Readings in Information Retrieval*. Morgan Kaufmann, pp. 47--59.

Delling, D., Gaertler, M., Görke, R., Nikoloski, Z., Wagner, D. (2006). How to evaluate clustering techniques. Technical Report 2006-24, ITI Wagner, Faculty of Informatics, Universität Karlsruhe (TH).

Fellegi, I. P., Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, Vol. 64, No. 328, pp. 1183--1210.

Luo, X. (2005). On coreference resolution performance metrics. In *Proc. of HLT-EMNLP*, pp 25--32.

Mann, G. S., Yarowsky, D. (2003) Unsupervised personal name disambiguation. In *Proceedings of CoNLL-7*, pp. 33--40.

Menestrina, D., Benjelloun, O., Garcia-Molina, H. (2006). Generic entity resolution with data confidences. In *First International VLDB Workshop on Clean Databases*. Seoul, Korea.

Spock Team (2008). The Spock Challenge. http://challenge.spock.com/ (Retrieved February 5.)

Vilain, M. Burger, J. Aberdeen, J. Connolly, D., Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference* (MUC6). Morgan Kaufmann, pp. 45--52.

Voorhees, E. M. (2001). The Philosophy of Information Retrieval Evaluation. In *Cross-Language Evaluation Forum*, pp. 355--370.

Creating a Gold Standard for Person Cross-Document Coreference Resolution in Italian News

Luisa Bentivogli, Christian Girardi, Emanuele Pianta
Fondazione Bruno Kessler (FBK-irst)
Via Sommarive 18 - POVO, 38100 Trento - Italy
E-mail: bentivo@fbk.eu, cgirardi@fbk.eu, pianta@fbk.eu

Abstract

This paper presents work aimed at the realization of a gold standard for cross-document coreference resolution of person entities in a corpus of Italian news. The gold standard has been created selecting a number of person names occurring in Adige-500K, a corpus composed of all the news stories published by the local newspaper "L'Adige" from 1999 to 2006. The corpus consists of 535,000 news stories, for a total of around 200 million tokens.To sample the person names in the corpus, we identified two dimensions, corresponding to two phenomena we intended to study, namely (i) the fame of the person entities and (ii) the ambiguity of person names. The first version of the gold standard is composed of 209 person names corresponding to 709 entities, for a total of 43,704 annotated documents.

1. Introduction

Recent years have seen an increase in the demand of content-annotated resources for new tasks in Natural Language Processing, such as content extraction and coreference resolution. Content extraction refers to the extraction of entities (e.g. persons, locations, organizations) and of relations between entities (e.g. affiliation of a person to an organization), while coreference analysis is the process of determining whether or not different text portions refer to the same entity. Entity coreference can be found both within the same document (intra-document), and across different documents in a corpus (cross-document).

Various initiatives such as MUC, ACE, SemEval made large annotated resources available and introduced quantitative evaluation, allowing remarkable advances within the fields of intra- and cross-document coreference. However, while such efforts are stimulating research for the English language, little has been done for other languages, where these kinds of resources are still lacking.

This paper constitutes our contribution to the field of cross-document coreference resolution, presenting a work aimed at the creation of a gold standard for cross-document coreference resolution of named person entities in Italian news.

This work has been carried out in the context of the OntoText (From Text to Knowledge for the Semantic Web) project, which has been funded by the Autonomous Province of Trento under the FUP-2004 research program. Based on the philosophy of the Semantic Web, OntoText exploits text processing and automatic reasoning technologies to extract knowledge from texts and organize it conceptually in an ontology. The new OntoText technologies have been applied and tested on the Italian corpus *Adige-500K*, which contains the news stories published by the local newspaper "L'Adige" from 1999 to 2006. The corpus consists of 535,000 news stories, for a total of around 200 million tokens. One of the main outcomes of the project is represented by the OntoText Portal, which provides an integrated access to the information automatically extracted from Adige-500K. Differently from common text-based search engines, the OntoText Portal directly accesses the concepts and entities of the ontology and presents the user with structured information instead of mere portions of texts. As specifically regards entities of type PERSON, when an OntoText Portal user types a person name as a query, he/she is presented with a set of clusters, where each cluster represents a specific entity and is assumed to contain all and only the newspaper articles referring to such entity.

One of the first uses of the gold standard is the evaluation of the coreference algorithm in charge of clustering the newspaper articles of the Adige-500K corpus according to the query of the OntoText Portal user.

The paper is structured as follows. Section 2 reports on other existing resources for cross-document coreference evaluation. Section 3 describes in detail the creation of the gold standard: its design, the annotation process, and all the data about its composition up to now. Section 4 presents the web interface specifically developed for the cross-document coreference annotation task. Finally, Section 5 draws some conclusions and explains future work.

2. Related Work

While intra-document coreference is a long dating and established area of research (e.g. anaphora resolution), the work on cross-document coreference resolution[1] began more recently. Bagga and Baldwin (1998) created the first reference data set for benchmarking cross-document coreference results, the *John Smith Corpus*, composed of 197 articles from the New York Times containing the name “John Smith”. The John Smith corpus allows for evaluating only a subset of the cross-document entity coreference functionality as documents containing name variations of “John Smith” are not included.

The field has seen a rapidly growing interest (Mann and Yarowsky 2003, Gooi and Allan 2004, Blume 2005, Bollegala et al. 2006), however the algorithms for coreference resolution were generally evaluated on very

[1] In the literature, this task is also referred to, with slight different meanings, as cross-document/interdocument/global coreference resolution, entity disambiguation, identity resolution.

few names in small corpora, or on artificial corpora, or through a posteriori control.

The first large size gold standard, which up to now represents the state of the art, has been created for the first cross-document coreference evaluation campaign, namely the SemEval-2007 Web People Search task (Artiles et al., 2007). The Web People Search corpus includes documents about 79 complete person names (first name and last name) corresponding to 1,882 entities mentioned in about 7,900 web pages (the 100 top results for a person name query to the Yahoo! search engine). The Web People Search corpus does not include documents with name variants and thus does not allow for name variation evaluation.

The resource that is most similar to our work is the forthcoming evaluation corpus of the ACE 2008 Global Entity Detection and Disambiguation task (ACE 2008), whose guidelines represent the standard to which we adhere. The ACE 2008 task consists in cross-document entity disambiguation, limited to documents in which entities are mentioned by name, be it the exact name or a name variant (e.g. long and short form of the name, variant spellings, misspellings, transliterations, aliases, and nicknames). According to the task, in the gold standard only coreference between named entities will be annotated.

The ACE 2008 corpus contains English and Arabic texts and will be composed of 10,000 documents per language. Only a subset of the whole corpus will be annotated for cross-document coreference.

As said above, all the resources available to the community up to now are for English. The only Italian resource annotated with cross-document coreference is the Italian Content Annotation Bank (I-CAB). I-CAB (Magnini et al., 2006) consists of 525 news documents taken from the local newspaper "L'Adige" for a total of around 182,500 words. The selected news stories belong to four different days (September, 7th and 8th 2004 and October, 7th and 8th 2004). The annotation of I-CAB has been carried out manually within the OntoText project, following the ACE annotation guidelines for the Entity Detection task, slightly modified to cope with the different morpho-syntactic characteristics of Italian. I-CAB is annotated with temporal expressions and with four types of entities, namely PERSON, ORGANIZATION, GEO-POLITICAL and LOCATION. Manual intra-document coreference has been carried out for all the annotated entities, with Callisto. Moreover, for PERSON and LOCATION entities also cross-document coreference has been carried out.

However, I-CAB is not suitable enough for evaluating cross-document coreference resolution as the newspaper articles have been chosen within a short time-span where very few different mentions of the same entity are found.

3. Creating the Gold Standard

Cross-document coreference of a person entity occurs when the same person is mentioned in more than one text source. It can be defined as a clustering problem, which in principle requires the clustering of name occurrences in a corpus according to the persons they refer to. In this work, as in SemEval, we consider clusters of documents containing the name occurrences. Cross-document coreference involves two problematic aspects, namely (i) to resolve ambiguities between people having the same name (i.e. when identical mentions refer to distinct persons) and, conversely, (ii) to recognize when different names refer to the same person.

The gold standard described in this paper addresses the annotation of cross-document coreference of named person entities in an Italian newspaper corpus.

The documents of the gold standard are selected from the Italian Adige-500K corpus. Given the number of documents in the corpus (more than 500,000) and the time-span covered (7 years), we think that Adige-500K is suitable for evaluation (and possibly training) of cross-document coreference, allowing for a great variety of name mentions and for entities to occur in a lot of documents.

Following the ACE 2008 guidelines, the annotation is limited to documents in which the entities are mentioned by name[2]. Different kinds of name variants are considered, such as complete names (Paolo Rossi), abbreviations (P. Rossi, Paolo R.), first names only (Paolo), last names only (Rossi), nicknames (Pablito), and also misspellings (Paalo Rossi) and journalist errors in reporting the correct name of the entity (Carlo Rossi instead of Paolo Rossi).

A representative number of names occurring in the Adige-500K corpus have been selected as seeds for the creation of the gold standard (*Seed Names*). Among all the possible name variants, we decided that a Seed Name is always a complete name, i.e. a pair First Name-Last Name (e.g. Paolo Rossi, Isabella Bossi Fedrigotti, Diego Armando Maradona)[3].

In order to select the Seed Names to be annotated, two main criteria have been adopted, corresponding to two phenomena that we intended to study. These criteria are discussed in the next section.

3.1. Gold Standard Design Criteria

The first issue to be addressed when creating the gold standard is how to sample the Seed Names.

Two dimensions have been selected, namely (i) the fame of the entities and (ii) the ambiguity of the Seed Names. The first dimension, which refers to the entity level, is strictly related to the context of application of the OntoText project, within which the gold standard has been created; the ambiguity dimension, which refers to the Seed Name level, is inherent in the cross-document coreference resolution task.

3.1.1. Entity Fame

Within the OntoText project we stressed the importance of the application context of the technologies developed, i.e. the OntoText Portal. We want to choose Seed Names

[2] In terms of the ACE categories, the entities considered are of type "PER", subtype "Individual" and class "SPC" (i.e. a particular, specific and unique entity) while the mention type is "NAM" (i.e. a proper name reference to the entity).

[3]In a complete OntoText application scenario, the Seed Names should represent all the possible user's query, i.e. all the name variant types. In the current version of the gold standard, we have introduced the restriction that the Seed Name is always a complete name but we are planning to add new Seed Names corresponding to other name variants.

which are representative of the OntoText Portal user queries.
We do not have data about actual user queries yet. However, we foresee that fame will be an important criterion to classify user queries. A great part of the user queries will be related to famous persons (which thus need to be adequately sampled in the corpus); however the user is likely to be asking information also about persons he/she knows, but who are not famous. For this reason we decided to include in the gold standard people belonging to five fame categories:

- Not famous
- Quite famous at the regional level
- Quite famous at the national level
- Very famous at the regional level
- Very famous at the national level

The distinction between the regional and the national level comes from the fact that the newspaper "L'Adige" contains both a national and a local section.

3.1.2. Name Ambiguity

The difficulty of the automatic coreference task varies on the basis of the ambiguity of the Seed Name: the more ambiguous the Seed Name, the more difficult is to disambiguate it. We want to study three different ambiguity scenarios:

- Low ambiguity
- Medium ambiguity
- High ambiguity

Summing up, we wanted the corpus to be structured along the two orthogonal variables of entity fame and Seed Name ambiguity. The original design of the gold standard is shown in Table 1, which partitions the expected set of entities of the gold standard in 15 cells; each cell is illustrated by the name of a sample entity.

	Not famous	**Quite famous regional**	**Quite famous national**	**Very famous regional**	**Very famous national**
Very ambiguous	Paolo Rossi	Elena Marino	Paolo Rossi		Paolo Rossi
Ambiguous	Franco Marini	Vittorio Colombo	Giovanna Marini		Franco Marini
Not ambiguous	Bruno Kessler	Dante Clauser	Marta Russo	Bruno Kessler	Umberto Eco

Table 1. Original design of the gold standard

In the original design, each cell in the grid was to be populated with entities, randomly selected from the Adige-500K corpus. However, to be able to use the standard evaluation techniques which are based on groups of entities carrying the same name (or variants of it), we decided that when we select an entity carrying a certain Seed Name for one cell, we also consider in the gold standard all other entities carrying the same Seed Name. Each time a given entity is introduced in the gold standard, also the other entities carrying the same name are introduced. This makes a full balancing of the gold standard difficult to achieve.
Moreover, as hinted by the two empty cells in Table 1, some cells are intrinsically scarcely populated, namely those containing entities very famous at the regional level and carrying ambiguous names. This is explained by the fact that, in general, there are much more unambiguous names than ambiguous ones. Even rarer are the ambiguous names which occur in the corpus and refer to famous persons. All these constraints make the task of populating the "famous" class difficult, especially in a regional context, which is more restricted than the national one.
Thus we gave up the idea of a full balancing of the two variables (which implies selecting the same number of entities for each cell), and we decided to have all the classes of ambiguity and all the classes of fame populated with a minimum number of entities, which has been fixed at 30.

3.2. Selecting Seed Names

Starting from the list of all the 592,000 Named Entities of type PERSON automatically recognized in Adige-500K, we created a list of gold standard candidates by selecting those Named Entities (i) composed of at least two words, (ii) occurring at least in five different newspaper articles, and (iii) occurring in no more than 1,000 newspaper articles.
The first constraint is necessary in order to obtain a complete Seed Name, which is composed of first name and last name. The second constraint has been adopted to obtain entities interesting from the point of view of the cross-document coreference. The third was adopted for the practical reason that manually annotating more than 1,000 documents for one single Seed Name is too time expensive and error-prone.
From the resulting list of 79,000 gold standard candidates, we randomly picked up Seed Names until we found those satisfying our sampling criteria. At the end of the process, we had selected 209 Seed Names. The rules followed in the selection are described in the next sections.

3.2.1. Entity Fame

As regards the entity fame dimension, the first problem to face was how to evaluate the fame of a given entity.
To this purpose, we selected a pool of people of different ages and we asked them whether they had heard about some proposed entities, identified by a complete name and a short description. Then, we used the answers to classify the entities in the five categories described above.
Table 2 shows the distribution of the 209 selected Seed Names over the five fame categories. It is important to notice that at this preliminary stage of the gold standard creation we could work only at the Seed Name level. This is due to the fact that the knowledge about the actual different entities corresponding to a given Seed Name is not available a priori but only at the end of the coreference resolution process. Thus, the "famous" cells of Table 2 contain Seed Names for which we knew that there was at least one famous entity, whereas the "not famous" cell contains Seed Names not referring to any famous entity. Section 4.1. reports data about the actual

entity fame, obtained after manual Seed Name disambiguation.

Not famous	Quite famous Regional	Quite famous National	Very famous Regional	Very famous National
59	42	38	23	47

Table 2. Population of the fame categories

As it can be seen in Table 2, the category of very famous people at the regional level is populated with only 23 Seed Names instead of 30. This is due to the fact that, given the nature of the newspaper, almost all the people which are very famous at the regional level carry a name which belongs to the group of the top frequency names having more than 1,000 occurrences in the corpus, which were previously excluded from the gold standard.

3.2.2. Name Ambiguity

In order to evaluate the ambiguity of a Seed Name, we resorted to an external source (see Artiles et al., 2007). The source used is PagineBianche, the Italian telephone directory. We exploited the information related to the number of subscribers having the same name to create three ambiguity classes according to the thresholds reported in Table 3. Then the three classes were as much as possible equally populated. Table 3 also reports how the 209 Seed Names selected were grouped with respect to PagineBianche. The class of unambiguous Seed Names is much more populated than the other two classes. This is due to the fact that almost all the Seed Names selected in order to populate the classes of famous entities (first sampling criterion) belong to the class of unambiguous Seed Names.

Ambiguity	Number of subscribers	Selected Seed Names
Not ambiguous	0-99	121
Ambiguous	100-199	42
Very ambiguous	200+	46

Table 3. Population of the ambiguity classes

PagineBianche is the only large scale representation of the Italian population that we could find. Unfortunately, such representation is not totally accurate. The subscribers of PagineBianche are usually adult males permanently living with their family. Young people who only own mobile phones are not present in PagineBianche and the same is for the majority of women because the PagineBianche subscriber is usually their male partner. We decided to normalize the occurrences of women names multiplying by five the number of female names found in PagineBianche.

3.2.3. Number of Documents

Another dimension of the corpus that we tried to keep under control (by a posteriori check) is the number of documents where a Seed Name occurs, as this can have an influence on the difficulty of the coreference resolution task. The frequency range fixed a priori goes from 5 to 1,000 occurrences of a given Seed Name in different newspaper articles. The number of documents containing the selected Seed Names cover most of this frequency range, with a minimum of 5 documents per Seed Name and a maximum of 893.

These ranges represent an approximation of the final number of documents associated to each Seed Name, as in this phase of the project the data about name variation of the Seed Names are not available yet.

The expected minimum size of the gold standard amounts to 32,582 Adige-500K documents, that is the number of documents containing a Seed Name mention.

Information about the correlation between the criteria according to which the gold standard has been modelled and the actual corpus data, which cannot be known a priori but only once the annotation has been carried out, will be given in Section 4.

3.3. The Annotation Process

To carry out the gold standard annotation, five annotators were selected and trained.

Given a certain Seed Name, the annotators have to disambiguate all the entities carrying that name and, for each entity, to find all the newspaper articles in which such entity is mentioned, both with its Seed Name and with all its possible name variants.

In this phase of the project, the annotators annotate the documents in which an entity is mentioned (in all its possible variants), but they do not annotate the single mentions of the entity within the documents.

In order to find all the possible name variants, the annotators can rely on a "lexicographer toolbox" (Giuliano, 2002) containing both concordances and collocations for the Adige-500K corpus. The toolbox turned out to be especially useful to find short forms of the names and misspellings.

The name variants found are used (together with any contextual word sequence identifying the entity) to create queries to the corpus, queries aimed at finding all the documents referring to the entity under consideration.

At the end of the annotation process, for each entity the result is (i) the identification of the documents referring to the entity and (ii) the creation of a list of its name variants.

According to the annotation guidelines, annotators are requested to take into consideration only entity mentions of type "proper name". In some cases the documents should not be annotated because:

- The entity is not mentioned with a proper name. This is the case of entity descriptions (e.g. "Il Sindaco di Trento"/ "The Mayor of Trento" for the entity "Alberto Pacher")
- The Seed Name refers to a non-person entity (e.g. organizations, streets, buildings named after a person, person names within titles of books, songs, etc.)
- The proper name refers to the author of the newspaper article.

In the case of non-informative documents, they are assigned to a "catch all" cluster. This happens for those documents containing only lists of names without any kind of further information.

As for the type of document annotation, the annotation can be marked as "not sure" in those cases where the

annotator is not sure if a document is referring to a specific entity or not.
In those cases where the same document refers to more than one entity carrying the same Seed Name, that document is assigned to all the different entities it refers to.
Regarding the information associated to the entities, for each entity the annotators report (i) the real, anagraphic, name of the entity (based on the annotator knowledge and/or other external resources, (ii) its group name, i.e. the Seed Name, (iii) a free description of the entity, and (iv) any kind of comment it could be necessary.
Another important characteristic of the annotation is the possibility of marking an entity as "similar to" another. This flag is used when the annotator is not sure if two (or more) apparently different entities are the same or not and it can be useful also for evaluation purposes as it allows to change the granularity of the gold standard clustering (more fine-grained if the entities are kept separate or more coarse-grained if they are kept together). All these different kinds of information are annotated into the gold standard through an annotation interface created for that purpose. The interface is described in detail in Section 5.

4. The Gold Standard

The current version of the gold standard is composed as shown in Table 4.

Seed Names	Entities	Documents
209	709	43,704

Table 4. Composition of the gold standard

The next sections report a posteriori data about the actual gold standard corpus after the application of a priori criteria, i.e. (i) the fame of *all* the entities, (ii) the corpus ambiguity of the Seed Names, and (iii) the total number of articles referring to a given Seed Name or a given entity.

4.1. Entity Fame

As already noticed in Section 3.2.1, the entity fame dimension turned out to be difficult to represent, due to the high number of occurrences of famous entities in the corpus. As a matter of fact, the goal of populating each "fame group" with a minimum number of 30 entities each has not been completely reached, as the class of "very famous at the regional level" contains only 24 entities.

Entity fame level	Number of entities
Not famous	542
Quite famous - regional	51
Quite famous - national	44
Very famous - regional	24
Very famous - national	48
Total entities	709

Table 5. composition of the corpus with respect to the entity fame dimension

As expected, we can see in Table 5 that the number of non famous entities is very high in comparison with famous entities. This is due to the fact that given a Seed Name referring to one famous entity, the same Seed Name often refers also to a number of non famous entities.

4.2. Seed Name Ambiguity

The average Seed Name ambiguity in the corpus amounts to 3.39. In order to verify if the PagineBianche can be considered a reliable source for assessing Seed Name ambiguity and if the thresholds we chose are adequate, we calculated the corpus ambiguity of the Seed Names selected from the three ambiguity ranges of PagineBianche. Table 6 shows that there is a correlation between the PagineBianche ambiguity ranges and the actual corpus ambiguity.

PagineBianche ambiguity ranges	Seed Names	Number of Entities	Average corpus ambiguity
Low	121	256	2.12
Medium	42	154	3.67
High	46	299	6.50
All corpus	209	709	3.39

Table 6. Seed Names ambiguity in the corpus

4.3. Number of Documents and Name Variation

The number of documents per entity, after annotation, ranges from 1 document to 1,419, and is well distributed on the whole range.
Among the 32,582 documents containing the Seed Names, 6,637 were not annotated, as they refer to non-person entities or to the journalists who wrote the articles (see Section 3.3).
The total number of documents composing the current version of the gold standard amounts to 43,704, among which 25,945 contain the exact Seed Name and 17,759 contain only name variants.
As regards the different types of name variants occurring in the texts, data about how many name variant types can be found within the annotated documents are not available up to now because the intra-document coreference annotation has not been carried out yet.

5. The Web Annotation Interface

A multi-user web interface was specifically designed for the cross-document coreference annotation task.
The interface is composed of two pages, the *Entity Management Page* and the *Document Annotation Page*, illustrated in Appendix 1. The *Entity Management Page* (Figure 1) contains all information about entities. In the left hand side the *Entity Search* functionality can be found. This functionality allows the annotator to look up a specific entity, to retrieve the list of documents associated to it, and to select the entity for the work session.
In the right hand side of the page, the *Entity Record* and the *Work Session* can be found. The *Entity Record* contains several fields where the annotator inserts and

modifies (i) the real anagraphic name of the entity (e.g. Guido Giuseppe Rossi), (ii) the group name, corresponding to the Seed Name (e.g. Giuseppe Rossi), (iii) a short description characterizing the entity, (iv) the identifier of possible similar entities, (v) an entity fame indicator (according to the annotator's knowledge), and (vi) a comment with all useful information. Moreover, when necessary, the entity can be marked as "catch all" (see Section 3.3). The *Work Session*, on the bottom right side, contains all the entities created in correspondence with a given Seed Name and is used as entity repository during the document annotation process. In some cases it can happen that two different entities turn out to be the same. The "merge" button allows the annotator to merge the two entities without having to annotate the documents again.

The *Document Annotation Page* (Figure 2) has the same layout of the OntoText Portal. The annotator submits a query and obtains all the documents satisfying the query, together with the text snippet in which the query string occurs.

A scroll down menu is associated to each retrieved document, where the annotator can select the entity to which the document refers. The entities presented for annotation correspond to those inserted in the Work Session created by the annotator in the Entity Management Page. If the document snippets are not informative enough to individuate the correct entity, the annotator can also access the full article. If the document turns out to be really difficult to be assigned to an entity, the annotator can mark the annotation as "not sure" by clicking the button at the left of the scroll down menu.

When the results of an annotator query are displayed, all the documents already annotated according to the entities contained in the Work Session are highlighted and the annotator can decide if he/she prefers to see them in the page or to hide them.

6. Conclusion and Future Work

We presented work aimed at developing a gold standard for person cross-document coreference resolution. The first version contains 209 different names, 709 different entities, and more than 43,700 newspaper articles.

We think that such an extensive gold standard can help assess and advance the state of the art for cross-document entity coreference resolution. However, the sampling criteria followed to generate the gold standard, especially the suitability of the external source used to determine Seed Name ambiguity, the method of evaluation of the entity fame, and the balancing of these two dimensions, represent issues which are open to discussion.

Up to now, we have gathered the data necessary to calculate the inter-annotator agreement, which will refer to 20 Seed Names (10% of the total) selected from the different cells composing the gold standard. The annotation has been performed by two of the five annotators who worked at the gold standard.

As regards the metrics to be used to calculate intercoder agreement, different measures have been proposed in the literature in the last years, the most used for NLP tasks being the K measure. Recently, the suitability of the traditional K measure has been put under discussion (Artstein and Poesio, to appear). As regards our specific field, the main problems relate to the fact that (i) in a clustering task there is not a common and predefined set of categories (the different person entitities), and (ii) the distribution of the number of clusters and and their size is not homogeneus among the different Seed Names. We have not calculated inter-annotator agreement yet. However, as a preliminary assessment, we carried out the evaluation of the two manual annotations with the SemEval-2007 Web People Search scorer. The scorer relies on the standard clustering measures of Purity, Inverse Purity, and F-measure. Table 7 reports the results obtained for the annotators, which can be compared with the "All-in-One" baseline run on Adige-500K.

	Purity	Inverse Purity	F-measure
Annotators	0.92	0.90	0.91
Baseline	0.86	0.77	0.81

Table 7. Preliminary evaluation of inter-annotator agreement

Table 7 shows that the manual annotation outperforms the All-in One baseline, suggesting that our gold standard has been annotated with a good intercoder agreement. Annotator 1 created 88 entities and annotated 5,176 documents, while Annotator 2 created 103 entities and annotated 5,030 documents.

As further future work, we plan to carry out the annotation of the intra-document coreference using the name variants found during the cross-document annotation.

Both the design of the gold standard and the various kinds of information contained in the annotations allow a wide range of possible evaluations.

The partition of the gold standard in 15 classes, representing the different levels of entity fame and Seed Name ambiguity, allows for a more informative evaluation and analysis of systems performances.

Concerning the task to be evaluated, exact name and name variations are considered, thus covering the whole cross-document coreference spectrum. As regards the evaluation itself, it is possible to set the gold standard clustering granularity (grouping or maintaining separate entities marked as similar) and to assign different scores to documents marked as "not sure" for the cluster to which they have been linked.

The first usage of the corpus has been the evaluation of the OntoText coreference algorithm (Popescu and Magnini 2007, Popescu 2008). To this purpose, we exploited the SemEval scorer.

As regards other uses of the gold standard, when the intra-document coreference annotation will be performed, it will also be possible to evaluate this task. Finally, we envisage its usage within the next edition of EVALITA (Magnini and Cappelli, 2007), a new initiative devoted to the evaluation of Natural Language Processing tools for Italian.

7. Acknowledgements

This work has been supported by the Ontotext project, funded by the Autonomous Province of Trento under the FUP-2004 research program. Our sincere thanks go to Diego Barbon, Giampier Bondoni, Fabio Niccolini,

Giuliano Tomasini, and Matteo Tonini who contributed to the development of the gold standard.

8. References

ACE 2008 Evaluation Plan http://www.nist.gov/speech/tests/ace/2008/doc/ace08-evalplan.v1.1.pdf

Artiles, J., Gonzalo, J., Sekine, S. (2007). Establishing a benchmark for the Web People Search Task: The Semeval 2007 WePS Track. In *Proceedings of SemEval-2007 Workshop*, co-located with ACL 2007, Prague, CZ, 23-24 June 2007.

Artstein, R., Poesio, M. (to appear). Inter-Coder Agreement for Computational Linguistics. In *Computational Linguistics*, to appear.

Blume, M. (2005). Automatic Entity Disambiguation: Benefits to NER, Relation Extraction, Link Analysis, and Inference. In *Proceedings of the International Conference on Intelligence Analysis,* McLean, Virginia, USA, 2005.

Bollegala, D., Matsuo, Y., Ishizuka, M. (2006). Extracting Key Phrases to Disambiguate Personal Name Queries in Web Search. In *Proceedings of the ACL'06 Workshop on How Can Computational Linguistics Improve Information Retrieval?*

Giuliano, C. (2002) A Tool-Box for Lexicographer. In *Proceedings of the Tenth EURALEX Congress*, Denmark, Copenhagen, August 13-17, 2002.

Gooi, C. H., Allan, J. (2004). Cross-Document Coreference on a Large Scale Corpus. In *Proceedings of the Human Language Technology Conference (HLT-NAACL)*. Boston, Massachusetts, 2004.

Magnini, B., Cappelli, A. (2007). Evalita 2007: Evaluating Natural Language Tools for Italian. Special issue of *Intelligenza Artificiale (AI*IA)*.

Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Bartalesi-Lenzi, V., Sprugnoli, R. (2006). I-CAB: the Italian Content Annotation Bank. In *Proceedings of LREC 2006* - Genova, Italy, May 22-28, 2006.

Mann, G. S., Yarowsky, D. (2003). Unsupervised personal name disambiguation. In *Proceedings of CoNLL-7*, 2003.

Popescu, O., Magnini, B. (2007). Web People Search Using Name Entities. In *Proceedings of SemEval-2007 Workshop*, co-located with ACL 2007, Prague, CZ, 23-24 June 2007.

Popescu, O. (2008). *Ontological Constraints in a Statistical Framework for Person Cross Document Coreference*. PhD Thesis, University of Trento. FBK Technical Report, April 2008

Appendix 1: the Annotation Interface

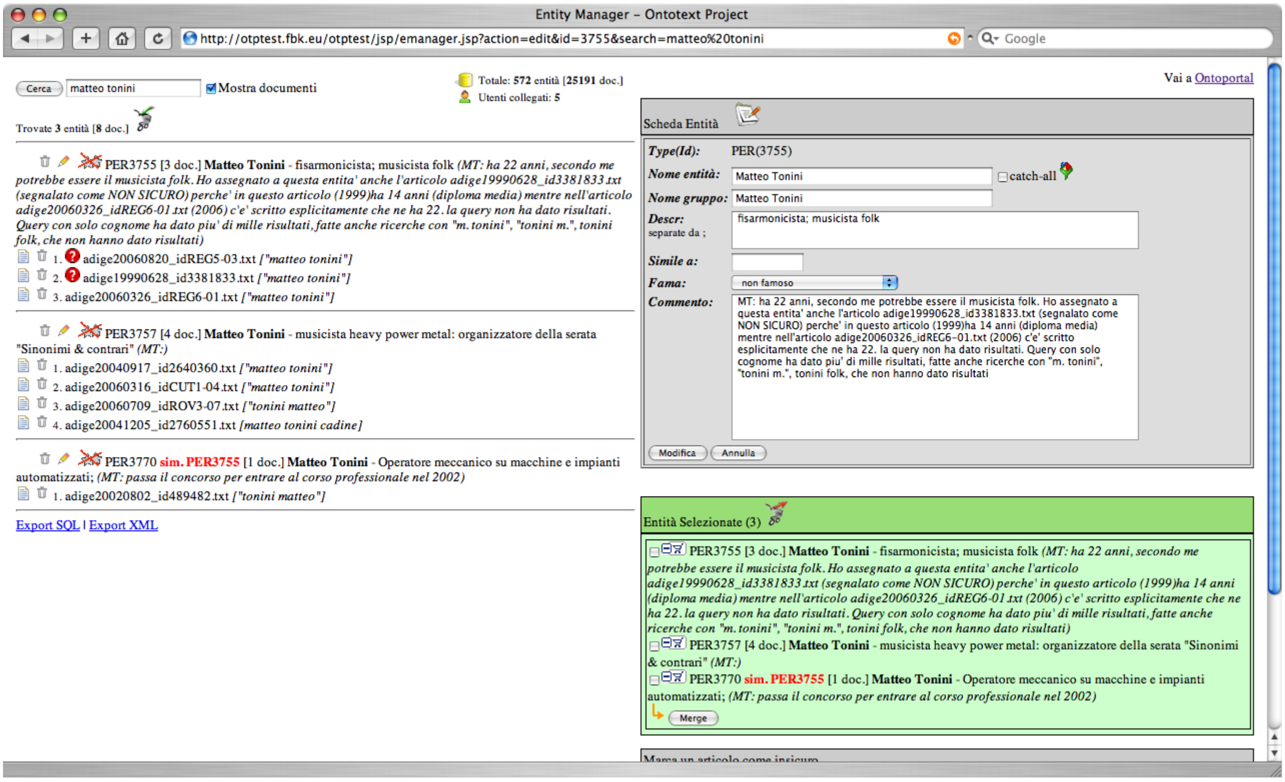


Figure 1. The Entity Management Page

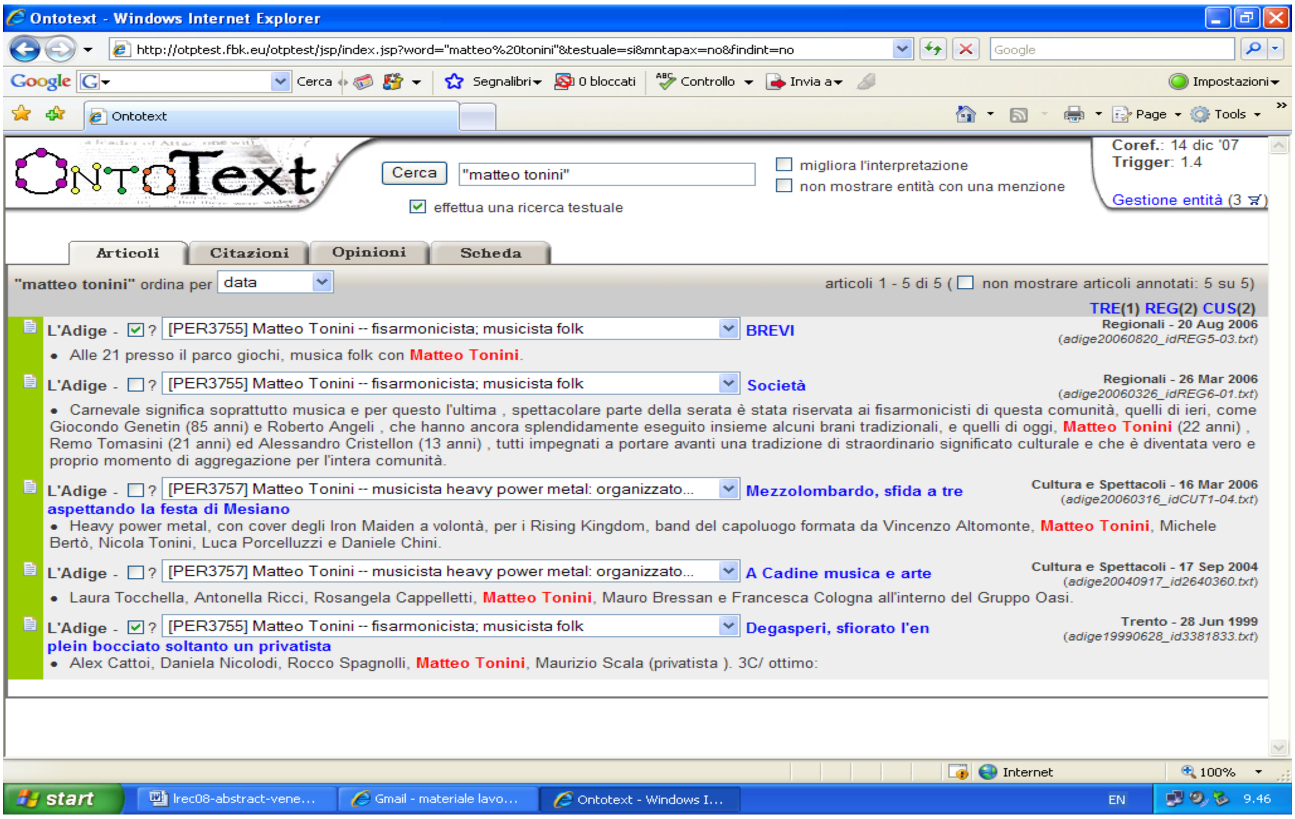


Figure 2. The Document Annotation Page

Methods for Evaluating Entity Disambiguation

Matthias Blume and Paul Kalmar
Fair Isaac Corporation
3661 Valley Centre Dr.
San Diego, CA 92130 USA
MatthiasBlume@FairIsaac.com, PaulKalmar@FairIsaac.com

Abstract

Within the last few years, interest in disambiguating mentions of entities found in plain text has surged. This paper describes best practices for evaluating entity resolution systems, including selecting representative evaluation data, machine-assisted generation of ground-truth assertions, metrics, and evaluation methods that do not require ground-truth data. The paper is written primarily from the perspective of disambiguating persons mentioned in plain text, but many of the methods are equally applicable to the disambiguation of other entity types and to sources of structured data other than information extraction from plain text.

1. Introduction

Entity disambiguation resolves the many-to-many correspondence between *mentions* (Mitchell, 2004) of entities in data records (such as text or transaction data) and unique real-world *entities*. This basic operation has been carried out in a variety of fields for decades (Newcombe, 1959) and has been referred to by a number of terms including *de-duplication*, *entity resolution*, *entity tracking*, *fuzzy matching*, *identity matching*, *merge/purge*, *object identification*, *record linkage*, *referential linking*, and *reference reconciliation*. Historically, most software for this purpose was tailored to a few *explicit* identification fields such as name, address, and telephone number.

Within the last few years, interest has surged in determining whether two snippets of text refer to the same entity (e.g. Bagga, 1998; Cucerzan, 2007). Entity disambiguation would allow a user to retrieve all records dealing with a particular entity, even if there are spelling variations in the entity's name, and without retrieving records corresponding to different entities with the same name. Entity disambiguation is essential for social network analysis and inference – given relations between "John Smith" and "Abdul Khan" and between "Abdul Khan" and "Kahuta Research Laboratories", it is impossible to determine whether "John Smith" and "Kahuta Research Laboratories" are related without determining whether the two mentions of "Abdul Kahn" refer to the same person. Finally, cross-language and cross-modality (speech-to-text) entity disambiguation permit translation and transcription to the correct name of that particular entity in the target language in ambiguous situations such as "Eric Smith" vs. "Erik Smyth".

Because *explicitly identifying attributes* are often not present in plain text, entity disambiguation here relies on utilizing *implicitly identifying information* (such as titles and relations). Importantly, many of the concepts necessary for disambiguating entities in plain text can be equally well applied to implicit ID in structured data (such as transactions) and to traditional explicit ID.

The accuracy of entity disambiguation systems varies tremendously depending on the type and amount of input data. With explicit ID, it easily exceeds 99%. With plain text (mostly implicit ID), the accuracy of determining whether "David Smith" (with or without spelling variations) in two documents refers to the same person is typically 90% to 95% (Blume, 2005). While good and in some cases exceeding human performance, fundamental improvements are still needed.

For example, when inferring a social network structure, each error causes an incorrect joining or separation of sub-networks, and the aggregation of errors results in qualitatively severe differences between the network model and the actual social network. Suppose that 10,000 documents mention Pakistan's president Pervez Musharraf. With agglomerative linking, a (typical) rate of 2% missed links yields 200 perceived entities. The vast majority of the documents will be associated with a single perceived entity, and most of the perceived entities will each be associated with a single document. Nonetheless, it would be disconcerting for an analyst to browse a portion of a social network and find dozens of Pervez Musharrafs that actually all correspond to a single person, and it can equally be dangerous to miss a single document with pertinent information about an entity of interest. It would be similarly frustrating to have 100 sub-networks incorrectly linked together by the incorrect merging of 100 persons named "Abdul Khan" (a very common name) into one.

Entity disambiguation has the potential to fundamentally improve applications ranging from Web search to intelligence analysis to the detection of money laundering. The performance of existing components is already very good, and subtle differences can potentially have large impact on downstream system performance. Thus, careful yet efficient evaluation is important for comparing as well as refining entity disambiguation systems.

The most interpretable but also the most laborious approach for evaluating performance is via the use of ground-truth data. This approach was used by Bagga and Baldwin in their seminal paper (Bagga, 1998), in the 2005

US government-sponsored Knowledge Discovery and Dissemination Challenge, in the Web People Search Task at the 4th International Workshop on Semantics Evaluation (Artiles, 2007), and in the 2008 ACE global entity detection and disambiguation task (NIST, 2008). Both the methods for selecting the evaluation data (section 2) and the evaluation metrics (section 3) have a tremendous impact on the quantitative and qualitative results of the evaluation. Utilizing a machine-assisted annotation approach (section 4) makes it possible to generate an order of magnitude more ground-truth data with the same effort, making a much more thorough evaluation possible. Finally, several approaches for evaluating entity disambiguation systems *without* the use of ground-truth data are described in section 5.

2. Selecting Representative Evaluation Data

Typically, it would be prohibitively time-consuming and expensive to annotate all entity mentions occurring in an evaluation corpus. However, manual specification of the correspondence between entity mentions and unique entities for a subset of the mentions is entirely feasible. If the mentions to be annotated are chosen to properly sample the characteristics of the full corpus, the system performance measured on the sample can also be extrapolated to infer the performance on the whole dataset.

In plain text data, it is more efficient in terms of the annotation effort to evaluate entity disambiguation against ground-truth information for *a few specific entities in all documents in which they appear* in a corpus (*longitudinal entity annotation*) vs. for *all entities in a few specific documents* (*transverse entity annotation*). For example, Bagga and Baldwin (Bagga, 1998) selected all documents containing the string "John Smith" (with some variations) and provided ground-truth correspondence to real-world entities for only those names, not any other names in the selected documents.

Entity disambiguation systems typically perform differently for entities with different characteristics, such as:

- Common names (e.g. "Li") vs. uncommon names (e.g. "Belitsina").
- Frequently mentioned persons such as famous people (in news) or prolific authors (in publication records).
- Various kinds of spelling variations.

Name spelling variations include:

- Reversal of given and family name.
- Optional name tokens, including middle names, titles, and suffixes (e.g. Junior).
- Abbreviations, e.g. middle initials.
- Short forms of names, such as "Rob" or "Bob" for "Robert".
- Nicknames and aliases, for example "Mahmoud Abbas, also known as Abu Mazen".
- Transliteration variations, e.g. "Mahmud" vs. "Mahmoud".
- Optional whitespace, hyphens, apostrophes, and diacritics.
- Capitalization variations.
- Typographical errors.
- Nominal mentions, e.g. "the President".

Different corpora will differ along different dimensions. Furthermore, names from some ethnic origins differ more or less along certain dimensions – a number of Chinese family names are extremely common, Japanese tend not to have a middle name, and Arabic person names tend to consist of many tokens, many of which are optional in discourse.

The set of target names should be selected to span variations in the above characteristics. This enables system performance evaluation in each of several dimensions and extrapolation to system performance on the entire corpus.

One possible method for creating a corpus with a target name is to query the Web for the target name. This method for creating an evaluation corpus was used in the Web People Search Task (Artiles, 2007). A problem with this type of corpus creation method is that it is not necessarily possible to extrapolate the results to other types of corpora as they have different characteristics. Using the target entity name in the query unnaturally increases the probability of the target names in the set of retrieved documents vs. the entire corpus (the Web). Furthermore, the set of retrieved documents is affected by the search algorithm and the frequency of occurrence of the target name. For example, when querying for less commonly mentioned names, some search engines retrieve a disproportionate number of genealogy Web pages.

Depending on the corpus and the use case, it is necessary to distinguish between *document level annotation* and *mention level annotation*. Document level annotation is the labeling of every document with the list of real world entities contained therein. Mention level annotation is the labeling of every mention of an entity within a document with the real world entity that it corresponds to. Document level annotation is simpler, as it does not require any internal annotation of the document. Mention level annotation requires finding every mention within every document and annotating them. For certain use cases, such as clustering web search results by real world person in each document, document level annotation is an appropriate choice. Mention level annotation, however, can always be reduced to document level annotation, and is more precise.

3. Metrics

There are two major classes of evaluation metrics for entity coreference: pairwise and clusterwise. Pairwise evaluation checks each assertion about the coreference status of a pair of documents or mentions. Clusterwise evaluations, on the other hand, treat entities as clusters of mentions or documents. These then map a cluster of documents or mentions to another cluster in a ground-truth set and evaluate the degree of match between clusters. Pairwise classification has the possibility of assigning a probability of coreference to each pair of mentions, or can assign a binary score of coreferent or non-coreferent. Although it is possible to assign a probability for each item of membership to a cluster, most systems that we are aware of only assign a binary score of membership or non-membership. The main advantage of pairwise evaluation is that it makes it easier to see where errors are occurring, and possibly the reasons why. The main advantage of clusterwise evaluation is that it describes a solution to the problem that better parallels the real world entities.

With document level annotation, it is possible to have a non-disjoint clustering. This would mean that it would be possible for a single document to simultaneously be a part of multiple clusters. It occurs when a document contains two or more distinct real world entities. The likelihood of this situation can be reduced by only selecting a specific query name to annotate, but is still an issue when a single page discusses two people with the same or similar names. To our knowledge, there are currently no established clusterwise evaluation metrics that are robust to non-disjoint clustering. Pairwise metrics could be used in some cases, but would be unable to distinguish clusters that are purely subsets of existent clusters.

We discuss the following established metrics: pairwise precision/recall/accuracy, mutual information, MUC precision/recall, B-cubed precision/recall, and purity/inverse purity. We also introduce a new metric, F-purity/F-inverse purity, which we created to address the shortcomings of the other metrics in use with non-disjoint clusterings[1]. All of these metrics work in the same manner for mentions or documents, but we will discuss them in terms of documents, as only documents can have non-disjoint clusterings.

3.1. Metrics for Evaluating Pairwise Assertions

3.1.1. Precision, Recall, Accuracy

Pairwise precision, recall, and accuracy can be used to assess the quality of system output providing a binary score of coreferent or non-coreferent. For every pair of documents, a correct positive (CP) or correct negative (CN) is defined when the pair is coreferent or non-coreferent respectively in both the system output and the ground-truth. False positives (FP) or false negatives (FN) are defined when the system and ground-truth disagree on whether the documents are coreferent.

Precision, recall, and accuracy are then defined in the standard way:

$$\text{Precision} = \text{CP}/(\text{CP}+\text{FP})$$

$$\text{Recall} = \text{CP}/(\text{CP}+\text{FN})$$

$$\text{Accuracy} = (\text{CP}+\text{CN})/(\text{CP}+\text{CN}+\text{FP}+\text{FN})$$

The result of each of these is a percentage, which can then be combined using F-measures. The most common way is to take the F-1 measure of Precision and Recall.

3.1.2. Mean Rank

One way to evaluate performance using pairwise analog-valued scores is to compute mean rank. Given m input records, one can arrange the scores into an mxm matrix such that 1.0 in index (i, j) indicates that records i and j correspond to the same entity. Since the scores are analog-valued, one can rank the elements of the upper triangle of the matrix (above the diagonal) by their score and compute the mean rank of the fields where it is known that they are not co-referent. A lower mean rank indicates better performance.

The area for concern with the rank-based scoring metric is that it requires pairwise (as opposed to entity cluster) assertions, and it requires analog values. The metric is not compatible with systems that produce binary entity assertions.

3.2. Metrics for Evaluating Entity "Cluster" Assertions

The assertion that documents a, b, c, and d mentioned "John Smith #1" is structurally very similar to the assertion that documents a, b, c, and d deal with "topic cluster #1". Thus, people have gravitated toward utilizing metrics developed for evaluating clustering systems to evaluate entity disambiguation systems. The key difference is that whereas clustering systems generally assign each document to a single topic, entity disambiguation systems may discover that a single document mentions "John Smith #1" *and* "John Smith #2".

In the discussion below, C represents the distinct *clusters* provided by the entity disambiguation system, L represents the distinct *labels* provided by the ground-truth data, and D represents the set of document-level entities. For those metrics with two formulae, only the formula for the precision metric is shown, with the recall metric derivable by switching all Cs and Ls.

[1] See http://web-people-search-task---semeval-2007.googlegroups.com/web/ClusterEvaluationMetrics.pdf for a side-by-side comparison of the formulae.

3.2.1. Mutual Information

A possible way of comparing two clusterings is to take the mutual information between the system clusters and the ground-truth. Mutual information is computed as the weighted average of the pointwise mutual information between each cluster in one set of clusterings and each cluster in a second clustering.

$$\sum_{c \in C} \sum_{l \in L} \frac{|c \cap l|}{|nEvents|} \log \frac{\frac{|c \cap l|}{|D|}}{\frac{|c|}{|D|} \frac{|l|}{|D|}}$$

$$= \frac{1}{|nEvents|} \sum_{c \in C} \sum_{l \in L} |c \cap l| \log \frac{|c \cap l||D|}{|c||l|}$$

There are several problems with using the Mutual Information metric. One major problem is that Mutual Information favors clusterings with uniform distributions. Mutual Information also favors outputs with high numbers of clusters. Unlike other metrics, it does not yield a score in the range of 0 to 1.

3.2.2. MUC Precision/Recall

MUC Precision and Recall are evaluation metrics that were devised for the Message Understanding Conference (Vilain, 1995). This precision metric calculates the number of clusters minus the number of missing links to the ground-truth labels, divided by the number of documents minus the number of clusters. The same process is repeated with system-generated clusters and ground-truth labels switched to compute recall.

$$\frac{\sum_{c \in C} |c| - |\{l \in L | c \cap l \neq \emptyset\}|}{\sum_{c \in C} |c| - 1}$$

$$= \frac{|D| - \sum_{c \in C} |\{l \in L | c \cap l \neq \emptyset\}|}{|D| - |C|}$$

Bagga and Baldwin discuss how this metric is not appropriate for entity disambiguation (Bagga, 1998). If a system or a ground-truth set indicates that no document is coreferent with another, this metric will cause a division by zero error. This situation can also occur if a clustering is non-disjoint and has a number of clusters greater than or equal to the number of clusters, yielding a negative percentage or division by zero error respectively. Therefore, it is not possible to use this metric for non-disjoint clusterings. This metric also completely ignores all singleton entities.

3.2.3. B-Cubed Precision/Recall

B-Cubed Precision and Recall are metrics created by Bagga and Baldwin (Bagga, 1998) to attempt to address the shortcomings of MUC Precision and Recall. One of the main problems that they attempt to counter is to create a metric that is strictly clusterwise. This metric is the precision computed and averaged for each document individually with its corresponding system-generated cluster and ground-truth label, reversing clusters and labels for recall. There is a distinct mapping for each document between system-generated cluster and ground-truth label.

$$\frac{1}{|D|} \sum_{d \in D} \sum_{c \in C | d \in c} \sum_{l \in L | d \in l} Precision(c, l)$$

$$= \frac{1}{|D|} \sum_{l \in L} \sum_{c \in C} \frac{|c \cap l|^2}{|l|}$$

This metric is not appropriate for non-disjoint clusterings, as it is possible to have incorrect/incomplete clusterings with at least equal precision and recall to what the ground-truth would get against itself. Leaving out an overlapping cluster would only benefit the score.

3.2.4. Purity/Inverse Purity

Purity and Inverse Purity are standard metrics for cluster comparison. The Purity metric maps each system-generated cluster to the ground-truth label which gives it the best precision, and then computes weighted average precision under this mapping, and reverses clusters and labels for inverse purity. There is a separate one way mapping between system-generated clusters and ground-truth labels and between ground-truth labels and system-generated clusters.

$$\frac{1}{|D|} \sum_{c \in C} \max_{l \in L} |c| * Precision(c, l)$$

$$= \frac{1}{|D|} \sum_{c \in C} \max_{l \in L} |c \cap l|$$

If documents are allowed to be part of multiple entities, or non-disjoint clusters, the following situation can happen.

Given that many entities in the clustering keys are singletons, appending a list of singleton entities with every possible document to the bottom of a clustering will only increase the score. A singleton entity in the response always has a purity of 100%, whereas a singleton entity will be ignored with regards to inverse purity as anything larger will always supersede it. Furthermore, appending one entity that contains all documents to the rest of the entities will always yield perfect inverse purity. Given that the distribution of cluster sizes is often zipfian, the purity lost from this entity is largely recovered from the sequence of singletons appended at the end.

Using this silly answer of a single entity occurring in all documents followed by a singleton entity per document typically yields a higher score than that of a serious disambiguation that has a few mislabeled entities.

Normally, there is a direct trade off between purity and inverse purity which would prevent a labeling like this from scoring well. However, given that an entity can be a part of multiple clusters and that we are taking a maximum score, this trade off is no longer present.

3.2.5. F-Purity/F-Inverse Purity

To deal with the evaluation of non-disjoint clusterings, we devised the F-Purity/F-Inverse Purity metrics. Similar to purity and inverse purity, these proposed metrics map each ground-truth cluster to the system-generated cluster which gives it the best harmonic mean of precision and recall, and then computes weighted average F-1 under this mapping. The difference between this metric and purity/inverse purity is that the maximum is taken of harmonic mean of precision and recall, rather than just the one being measured. This allows the system to measure the mapping to the best matching cluster, rather than just the one which is most precise or has the most recall. Although both precision and recall are measured in both F-Purity and F-Inverse Purity, it is still necessary to compute both to prevent clusters on either side from being uncounted.

$$\frac{1}{|D|}\sum_{c\in C}\max_{l\in L}\frac{2|c| * Precision(c,l) * Precision(l,c)}{Precision(c,l) + Precision(l,c)}$$

$$= \frac{1}{|D|}\sum_{c\in C}\max_{l\in L}\frac{2|c||c\cap l|}{|c|+|l|}$$

3.2.6. Comparison of Metrics

It is illustrative to compare the metrics defined above on a pair of examples:

Labels (ground-truth)	(a b c d) (d e) (f g h)
Clusters 1	(a b c d g) (e) (f h d)
Clusters 2	(a b c d e f g h) (a) (b) (c) (d) (e) (f) (g) (h)

Each letter a-h denotes a document, and each set of parentheses denotes a single entity occurring in that set of documents. Subjectively, Clusters1 and the Labels seem rather similar, whereas Clusters2 has no relationship to the Labels – it simply states that some entity occurs in all documents and each document contains some entity that occurs in no other document. Thus, a suitable metric should score Clusters1 better than Clusters2.

The results are listed in Table 1. Of mutual information, MUC precision+recall, B-cubed precision+recall, purity+inverse purity, and F purity+ F inverse purity, only F purity+F inverse purity has the desired characteristic. (Mean Rank is not included, as the example lists only binary assertions.)

Metric	Clusters 1	Clusters 2
Mutual Information	0.187	0.256
MUC Precision	0.166	NA
MUC Recall	0.333	NA
B-CUBED Precision	0.733	1.403
B-CUBED Recall	0.824	1.333
Pairwise Precision	0.538	0.357
Pairwise Recall	0.7	1
Pairwise Accuracy	0.826	0.357
F-Purity	0.79	0.585
F-Inverse Purity	0.765	0.626
Purity	0.778	0.75
Inverse purity	0.778	1

Table 1. The result of comparing Clusters1 against the Labels appears in the first column and the result of comparing Clusters2 against the Labels appears in the second column.

4. Machine-assisted Annotation

The process of generating ground-truth for evaluating entity disambiguation typically consists of a human annotator carefully examining multiple documents and external data sources (such as the Web) to (i) learn salient attributes of real-world entities and (ii) map the mentions in the documents to those real world entities based on similarities in the observed attributes. This can be laborious and tedious, especially when dealing with entities outside the annotator's subject matter expertise. The process can be greatly accelerated by automatically highlighting possibly salient attributes and automatically grouping documents with many shared attributes. Figure 1 shows an example of a single publication record that has been marked up in this fashion:

It is document **BT003** and would be presented to the annotator in sequence immediately after **BT001** and **BT002**. The entity name of interest is Kim S.-H., denoted by ****. The annotator has presumably already decided that **BT001** and **BT002** mention a single entity with that name, and now must determine whether **BT003** mentions the same person or a different person with this name. The **BT001** and **BT002** sprinkled through the record highlight that Hase T., Wada S., and Yoshimura R. are Kim S.-H.'s co-authors not just on the current paper but also on the paper described in record **BT001**. Furthermore, papers **BT001** and **BT002** appeared in the same journal (Transplantation Proceedings) and dealt with the same topic (86.6.4.1) as this paper. Based on these highlighted attributes, that annotator may conclude that **BT003** deals with the same Kim S.-H. as **BT001** and **BT002**, click the appropriate button in the user interface, and move on.

```
<DOC ID="2000282618">BT003
<DOCTITLE>Role of natural killer cells in the rejection of transplanted hearts in the
mouse model</DOCTITLE>
<DOCDATE>03 DEC 2004</DOCDATE>
<PERSON ID="625338" STD="p_j_chargui_p">Chargui J. BT008</PERSON>
<PERSON ID="625339" STD="p_t_hase_p">Hase T. BT001</PERSON>
<PERSON ID="625340" STD="p_a_izawa_p">Izawa A. BT023</PERSON>
<PERSON ID="625341" STD="p_sh_kim_p">Kim S.-H. ***</PERSON>
<PERSON ID="625342" STD="p_t_kishimoto_p">Kishimoto T. BT008 BT010</PERSON>
<PERSON ID="625343" STD="p_s_wada_p">Wada S. BT001 BT008 BT010</PERSON>
<PERSON ID="625344" STD="p_y_wantanabe_p">Wantanabe Y.</PERSON>
<PERSON ID="625345" STD="p_r_yoshimura_p">Yoshimura R. BT001 BT008 BT010</PERSON>
<LOCATION>Dr. T. Hase, Department of Urology, Osaka University School of Medicine, 1-4-3
Asahima-chi, Abreno-ku, Osaka 545-8585</LOCATION>
<LOCATION>Japan</LOCATION>
<SOURCE STD="TRPPA" ISSUE="32/7 (2080-2081)" YEAR="2000">Transplantation Proceedings AH001
BJ001 BT001 BT002 BT004 ... CV007 ... FQ001</SOURCE>
<CLASSIFICATION ID="86.6.4.1">IMMUNOLOGY AND INFECTIOUS DISEASES: TRANSPLANTATION
IMMUNOLOGY: Transplantation: Experimental AY001 BQ002 BQ012 BT001 BT002 BT004 ... CV007
DD002 EW001 FL003</CLASSIFICATION>
</DOC>
```

Figure 1. A single publication record that has been marked up automatically for machine-assisted entity disambiguation ground-truth annotation.

On one occasion, we hired two undergraduates to carry out ground-truth entity annotation in this fashion. They spent a total of 240 hours annotating 9,690 publication records in 100 name groups with a total of 704 distinct names. For example, the group for Kim S.H. also included Kim S.-H., Kim S.-H.M., Seok Hyung Kim, Seung Hyun Kim, Soo Hyun Kim, and Soon Ha Kim. The annotators determined that these records dealt with 4,218 distinct target entities, yielding 6,983 pairwise assertions that two records deal with the same entity and 1,908,771 pairwise assertions that two records deal with different entities in the same name group. The level of productivity (annotating 18 entities per hour on average) is remarkable. Our impression is that automatically highlighting salient attributes and automatically grouping documents with many shared attributes speeds up the annotation process by a factor of 10. A third-party assessment of a random subset of these assertions found that the annotators' error rate was about 3%.

The one cause of concern is that the algorithms for highlighting salient attributes and automatically grouping documents are not perfect, the annotator becomes sloppy and just agrees with the system-generated grouping, and this bias in the errors in the ground-truth data unfairly penalizes entity disambiguation systems that are based on different algorithms. This concern can be largely ameliorated by providing the ground-truth data to proponents of the various entity disambiguation systems post-evaluation. If each proponent argues for correction of ground-truth errors that conflict with his/her system output, the final outcome would be a nearly perfect ground-truth data set.

5. Evaluation Without Ground-truth Data

Generating a true ground-truth dataset is costly and time-consuming, the ground-truth data typically contains some errors, and the system performance may be markedly different on other datasets with different characteristics. Thus, it is natural to ask whether there is some way to evaluate or compare entity disambiguation performance without ground-truth data. This section describes three such methods. These mechanisms permit a broader coverage (larger number of labeled examples) than manually generating ground-truth data, but the results of such evaluations are less interpretable.

5.1. Correlation with Topic Clusters

Imagine running entity disambiguation on information extracted from a text corpus, where that information explicitly excludes document topic. In parallel, strip out all entity mentions from the corpus and cluster and the resulting documents using any clustering algorithm. Intuitively, one would expect to find some correlation between the entity "labels" and the topic cluster labels.

Given two entity disambiguation systems, one would expect the better system to produce greater correlation. To the extent that this is true, it is possible to compare entity disambiguation performance without ground-truth data!

As indicated in section 3.2, many comparison metrics are sensitive to the number of entities found by the system. Thus, it is beneficial to compare curves produced by the systems, with the number of entities produced on the x-axis. The higher curve indicates the better system. An example of such a plot is shown in Figure 2.

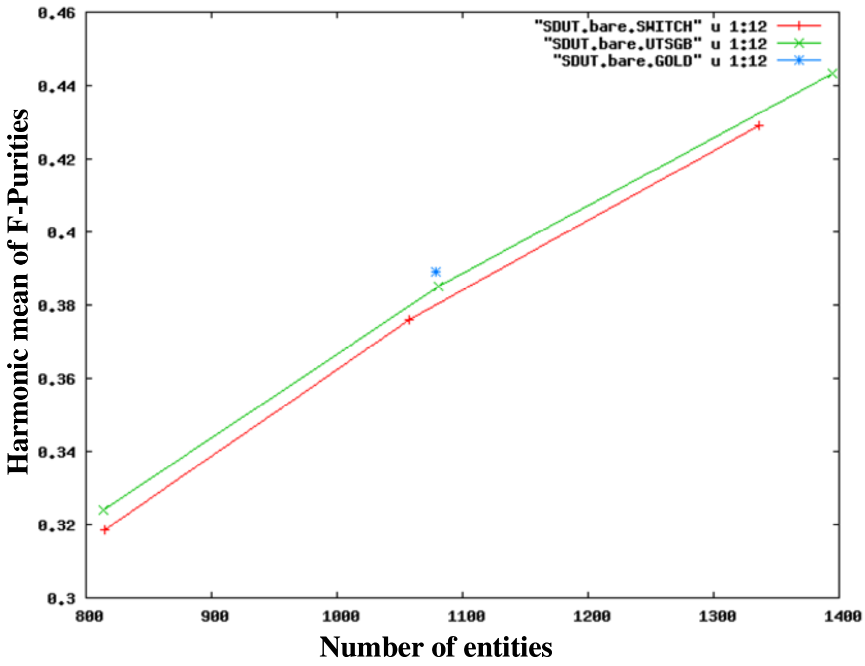


Figure 2. Harmonic mean of F-Purities between automatically determined entity labels and topic clusters for two different system configurations (curves) and between a set of ground-truth entity labels and the same topic clusters.

Even with varying thresholds for matching entities, it can be seen that one system consistently outperforms the other. These types of plots are useful for deciding general weights for various features of the disambiguation systems. They are not usable, however, for comparing disambiguation systems across separate teams as a team's feature selection might overlap with the data used to detect topics.

5.2. Name Truncation and Name Swapping

A set of methods for evaluating entity disambiguation systems without ground-truth data is based on the principle of stripping information out of a corpus prior to feeding the data into the entity disambiguation system. The system is handicapped because certain information is missing. Subsequently, it is possible to determine what set of assertions made by the system is incompatible with the hidden information.

For example, each string "David Jones" in the corpus can be replaced by "John Smith". Any system assertion that an altered mention (originally "David Jones") corresponds to the same entity as an unaltered mention (originally "John Smith") can be assumed to be incorrect. Similarly, any middle names and middle initials can be stripped, and any system assertion that confounds two mentions with different middle names can be assumed to be incorrect. This then could be evaluated using the mean rank metric described in Section 3.1.2, as was done in task ER1b of the 2005 Knowledge Discovery and Dissemination Challenge.

One area of concern with this evaluation method is that the process of substituting names creates synthetic data rather than natural data. Thus, the performance of a system on this task may or may not reflect the performance on real-world data, with the entity distributions and disambiguation challenges of the real world. Furthermore, the process of replacing names in the corpus changes the context information that is ultimately used to carry out the entity disambiguation. Thus, certain system errors may be attributable to the nature of the name substitutions that were carried out. Finally, some systems may latch on to inconsistencies in how the replacements were carried out and utilize such artifacts to attain artificially high disambiguation performance on the test records.

Also, it is only possible to detect certain false positive matches, not false negatives. Thus, a system that assigns each record to a distinct entity *might* be correct.

By operating on enough pairs of names, this methodology can be used to generate a greater number of tests than are feasible with a ground-truth dataset. This greater number of tests provides a greater statistical significance and numerical confidence in the system scores. However, using *only* this method would leave the above questions unanswered. Consequently, utilizing ground-truth is complementary to this method. The performance of systems on ground-truth data should correlate to that of the performance on name truncation data, and examining any discrepancies may lead to a better understanding of the entity disambiguation tasks and systems.

5.3. User-tagged Data

Another possible source of data for evaluating entity disambiguation without ground-truth is user-tagged data such as Wikipedia and some social network sites on the

Web. User-tagged data differs from true ground-truth data, as the intended use is very different and the annotation is often much less clean. Coreference information is provided by many users, rather than a small set of trained annotators and there are no checks for inter-annotator agreement. There is also no guarantee that the tags refer directly to coreference information and not to a larger containing entity.

Wikipedia is a large internet encyclopedia with pages annotated with links to other articles. If a link refers to an entity, theoretically all other mentions sharing that link will be coreferent, and all those not sharing that link will not. In addition, there are manually assigned category tags which describe the topic of each document, disambiguation pages which discriminate between different mentions with similar names, and list pages which describe in bulk the type of certain entities. Various papers have made use of Wikipedia as a ground-truth corpus for entity disambiguation, such as Bunescu (2006) and Cucerzan (2007). Both of these papers focus mainly on named entity discrimination (labeling an entity as a member of a previously defined set of labels), rather than disambiguation.

Although this creates an effective ground-truth corpus for this type of data, it is unclear how results on this type of corpus will apply to other types of corpora which have different characteristics. The papers which have used this corpus have used many corpus specific features such as category and link graphs. Also, encyclopedic data often contains articles which are strictly about a specific entity rather than discussing multiple entities at once.

The advantage of using this type of data as a standard for comparison is that it provides a large amount of data that is more accurate than name truncation or swapping and cheaper to produce in bulk than manually tagged ground-truth. The individual merges in user-tagged data are more easily read by a human than other artificially constructed ground-truths, and the reasoning behind a particular merge can be more easily understood.

6. Entity Types Other Than PERSON

While the above description has focused on the disambiguation of *person* entities, many of the concepts and methods are equally applicable to other entity types such as *organizations*, *locations*, *accounts*, *households*, or *vehicles* . Three potential differences are generic vs. specific entities, non-atomic entities, and entities from stable sets.

It is possible to disambiguate specific real-world items such as *the Toyota Prius with vehicle identification number 123456789* vs. *the Toyota Camry with VIN 987654321*. Each has particular *attributes* such as color, owner, license plate number, and location at any particular point in time. In contrast, entity disambiguation systems (or evaluation approaches) are generally not appropriate for distinguishing between generic entities such as *a Toyota Prius* versus *a Toyota Camry*.

Disambiguation of organizations (and evaluation thereof) is poorly defined in practice because organizations are not *atomic*. An organization may split into two new organizations, and two different organizations may merge into one. A department of a company is a reasonable *organization entity* that takes actions and definitely exists at some point in time, but a corporate reorganization may assign the departments' people and assets to different departments and/or companies. Organizations may own or legally control one another, such that the child organization is effectively a *part of* the parent organization. Various schemes exist for assigning IDs (such as DUNS and employer identification numbers) to organizations in the real world, but these IDs are in some ways more permanent than the underlying organizations. In practice, these differences can sometimes be ignored, especially if the data set of interest (e.g. world news) covers organizations at a level at which they are largely stable.

It is possible to compile a set of data records corresponding to most geopolitical entities (populated locations) that are likely to occur in any data set. Such a *gazetteer* could list coordinates, parent location, and population. The set of geopolitical entities is smaller, more stable, and has more readily accessible documentation than the set of persons. Thus, it is possible (and generally beneficial) to disambiguate location information extracted from text against a gazetteer, and this may be used to evaluate the disambiguation as well.

7. Conclusion

At the high performance levels provided by some existing entity disambiguation systems, careful evaluation is necessary both to quantify the level of performance and to test the impact of modifications to the technology in order to improve the systems. This evaluation is most reliable when carried out on ground-truth datasets. The evaluation metrics and the methods used to select the evaluation records can quantitatively and qualitatively change the outcome of the evaluation.

Most metrics for assessing the correspondence between system-assigned cluster labels and ground-truth cluster labels were developed under the assumption that each document is assigned to exactly one cluster. In entity disambiguation, it is entirely possible that a single document mentions two distinct entities with the same name. Many established clustering evaluation metrics are not appropriate for this scenario, for example rewarding the generation of spurious assertions. We introduce a new measure, F-purity + F-inverse purity, that does not suffer from these problems.

There are enormous differences among the characteristics of data sets to which entity disambiguation may be applied, such as name and address data, "clean" newswire, blogs (with user IDs), and Wikipedia (with lists, manual annotation, and meaningful links between documents). Thus, the results of each evaluation exercise are somewhat specific to the underlying corpus type. Utilizing machine-assisted annotation greatly speeds up the process of generating ground-truth data for a new corpus type.

Several methods exist for evaluating entity disambiguation systems without ground-truth data. However, these are less interpretable. It is possible to disambiguate entity types other than *person*. In some cases, it is sensible to disambiguate and evaluate against external data, such as a gazetteer. In principle, the disambiguation of non-atomic entities such as organizations is different from that of persons.

8. Acknowledgment

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

9. References

[Artiles, 2007] J. Artiles, J. Gonzalo, and S. Sekine, The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task. In Proceedings of Semeval 2007, Association for Computational Linguistics, 2007.

[Bagga, 1998] A. Bagga and B. Baldwin, "Entity-based Cross-document Coreferencing Using the Vector Space Model", *17th International Conference on Computational Linguistics (CoLing-ACL)*, p. 79-85, 1998.

[Blume, 2005] M. Blume, "Automatic Entity Disambiguation: Benefits to NER, Relation Extraction, Link Analysis, and Inference", *First International Conference on Intelligence Analysis*, 2005.

[Bunescu, 2006] R. Bunescu and M. Pasca, "Using Encyclopedic Knowledge for Named Entity Disambiguation", *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, p. 9-16, 2006.

[Cucerzan, 2007] S. Cucerzan, "Large-Scale Named Entity Disambiguation Based on Wikipedia Data", Conference on Empirical Methods in Natural Language Processing & Conference on Computational Natural Language Learning (EMNLP-CoNLL), 2007.

[Gooi, 2004] C. H. Gooi and J. Allan, "Cross-Document Coreference on a Large Scale Corpus", *Human Language Technology Conference (HLT-NAACL)*, p. 9-16, 2004.

[Huang, 2003] F. Huang, S. Vogel, and A. Waibel, "Automatic Extraction of Named Entity Translingual Equivalence Based on Multi-feature Cost Minimization", *ACL-03 Workshop on Multilingual and Mixed-language Named Entity Recognition*, p. 9-16, 2003.

[Kalashnikov, 2005] D. V. Kalashnikov and S. Mehrotra, "A Probabilistic Model for Entity Disambiguation Using Relationships", *SIAM International Conference on Data Mining (SDM)*, 2005.

[Kalmar, 2007] P. Kalmar and M. Blume, "Web Person Disambiguation Via Weighted Similarity of Entity Contexts", International Workshop on Semantic Evaluations (SemEval), 2007.

[Mann, 2003] G. S. Mann and D. Yarowsky, "Unsupervised Personal Name Disambiguation", *Conference on Computational Natural Language Learning (CoNLL)*, p. 33-40, 2003.

[Mihalcea, 2003] R. Mihalcea, "The Role of Non-Ambiguous Words in Natural Language Disambiguation", *Conference on Recent Advances in Natural Language Processing (RANLP)*, 2003.

[Mitchell, 2004] A. Mitchell, S. Strassel, P. Przybocki, J. K. Davis, G. Doddington, R. Grishman, A. Meyers, A. Brunstein, L. Ferro, and B. Sundheim, "Annotation Guidelines for Entity Detection and Tracking (EDT), Version 4.2.6", 2004.

[Newcombe, 1959] B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James, "Automatic Linkage of Vital Records", Science, vol. 130, pp. 954-959, 1959.

[NIST, 2008] National Institute of Standards and Technology., "Automatic Content Extraction 2008 Evaluation Plan", 2008.

[Ravin, 1999] Y. Ravin and Z. Kazi, "Is Hillary Rodham Clinton the President? Disambiguating Names across Documents", *ACL 1999 Workshop on Coreference and Its Applications*, p. 9-16, 1999.

[Vilain, 1995] M. Vilain et al., "A Model-Theoretic Coreference Scoring Scheme", *Procedings of the Sixth Message Understanding Conference (MUC-6)*, p.45-52, 1995.

Linking, mapping, and clustering entity records in information-based solutions for business and professional customers

Jack G. Conrad, Tonya Custis, Christopher Dozier, Terry Heinze, Marc Light, Sriharsha Veeramachaneni
Thomson Corporation
610 Opperman Drive, Saint Paul, MN 55123 USA
E-mail: marc.light@thomson.com

Abstract

This is a position paper that describes a number of use cases and their corresponding evaluation metrics. We discuss three types of resolution problems: linking entity mentions in text to records in a database, mapping records in one database to those in another database, and clustering records in a single database. The use cases arose at the Thomson Corporation and the systems developed support a number of products.

1. Introduction

The aim of this paper is to provide the reader with an overview of the entity resolution tasks we have worked on, the methods we have employed, and the evaluations we have used.

To provide context for our discussion, it is useful to have some idea of what our company does: the Thomson Corporation provides information-based solutions for lawyers, business people, nurses, doctors, scientists, and other professionals. Many of these solutions involve textual sources in combination with more structured sources such as databases of numeric and nominal information. Both the text and the databases contain information about entities ranging in type from genes to cities. Part of the "intelligent information" that Thomson products use is the mapping, and clustering of entity records along with linking of these records to text mentions.

Historically this mapping, clustering, and linking has been done manually. However, increasingly, automated systems are being used. In some cases, automated systems assist humans, improving their accuracy and efficiency. In other cases, the accuracy of the automated systems is sufficient alone. Our department, Thomson Research and Development, has been involved in such work and has developed a number of automated systems including systems that support products such as Westlaw Profiler (http://west.thomson.com/westlaw/profiler/), Westlaw Medical Litigator (http://west.thomson.com/westlaw/litigator/medical.aspx), and West's Monitor Suite (http://www.firm360.com/). In addition to working in the legal domain, in recent years, we have worked on systems for Thomson Financial, Thomson Scientific, and Thomson Healthcare.

The remainder of the paper is structured as follows. First, we discuss tasks of linking entity mentions in text to records in a database. Next we discuss mapping records in one database to those in another database; such a task arises when two databases need to be merged. Finally we discuss clustering records in a single database; such a task arises when a database contains numerous records for the same entity but there is no explicit information denoting the relation. For each of these three general tasks, we describe our general approach and evaluation methods and then describe one or more case studies.

2. Linking entity mentions in text to records in a structured database

We have created a number of applications that are based on extracting named entities from text and attaching them to structured records in an entity database. The basic method consists of the following two steps. First we extract from the text the entity names of interest along with information that can be used as evidence for entity resolution. Then we place the extracted text segments into a structured record called a template record and attempt to resolve (link or match) the template record to a record in an entity database. The first step in this process is called the extraction phase. The second step is called the entity resolution phase. We will only discuss the resolution phase here.

The entity resolution phase is based on record linkage techniques. The entity resolution phase can be separated into two phases: blocking and matching. In the blocking phase, we use some element of the extracted person name to read a subset of the records from the database likely to contain any existing database record matching the extracted person name. A typical blocking key might consist of all or part of a person's last name. Blocking is necessary because it is usually not computationally feasible to perform the full matching function on every database record for each extracted name. Blocking and its role in record linkage is further discussed in (Winkler, 1995) and (Baxter, et al., 2003). The second phase is matching and consists of comparing each database record in the block to the current template record and computing the likelihood that the template record and a given database record refer to the same person (i.e. match). The complexity of the resolution step is determined by the size and similarity of the entities in the database, the quality of the extracted data in the template record, the comprehensiveness of the database, and any contextual knowledge about the text that indicates whether the person names from the text are likely to belong the same set of people covered by the database.

For person names, the features we often use in our matching functions include the degree of match between the first, middle, and last name of the person and also include location information, appositive information indicating person's profession, and organization names with which the person is affiliated. We usually combine features to compute a match belief score using either naïve Bayes and support vector machine classifiers In some cases, we have used heuristic rules to combine the features to arrive at a decision. At this point, we do not have a principled process for deciding which type of classifier to use on a new problem.

We typically collect positive training data by asking editors to provide between 500 and 1000 manually matched examples chosen at random. We then collect very large amounts of negative training data automatically by pairing the template record from the positive data with all of the database records except the one identified as matching in the positive set.

After we learn our match function from the training data and compute match scores between every database record in the block and a given template record, the highest scoring database record is linked to the template record provided the match score exceeds a match threshold determined by the training data. If the highest scoring record falls below the match threshold, we check the score against a low threshold to determine if the template record is far enough away from all database records to warrant the creation of a new database record. If the match score falls below the low threshold, it is likely the template record refers to a new person and we therefore add it to the database. If the highest score falls between the match and low thresholds, we log the template record for manual review.

We usually measure the quality of our text to database linking systems using precision and recall as measured against a held out test set. We like to have a least 300 test records available, which often gives us a small enough confidence interval around the resulting precision and recall numbers. Our baselines start with a system that chooses at random from the returned block size. Thus, if the average block size is 2, then the first baseline would have an accuracy of 50% (precision 50%, recall 50%, and F-measure of 50%). Then, we provide progressively more intelligent baselines by using heuristics based on frequent high precision features, e.g., pick the record that has a location field closest in edit distance to the template field.

In the subsections that follow, we describe two specific applications that are based on the text-to-database record linkage methodology described above.

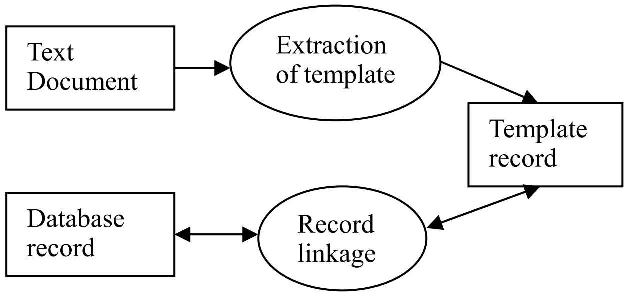


Figure 1: System diagram for linking entities in text to database records

2.1 Case study: linking legal professionals from caselaw documents to legal directories

In this task we extracted attorney, judge, and expert witness names from American caselaw, briefs, and professional journals. Then we attached these names to unique person records in a comprehensive database of U.S. legal professionals (Dozier & Haschart, 2000). By establishing these links, we are able to offer users the ability to browse through documents in which an individual is mentioned and to offer users the ability to jump to an individual's curriculum vitae from a name mentioned in text. New records are continually added to the person database when mined names do not match any individuals currently residing in the database.

A typical paragraph in caselaw that identifies the attorneys involved in a case is shown below.

> H. Patrick Weir, Jr., Lee Hagen Law Office, ltd., Fargo, N.D., Jeffrey J. Lowe, Gray & Ritter, P.C., St. Louis, MO, and Joseph P. Danis and John J. Carey, Carey & Danis, LLC, St. Louis, MO, for plaintiff and appellant.

Figure 2: Attorney paragraph

In the example paragraph, our system extracts and links H. Patrick Weir, Jr., Jeffrey J. Lowe, Joseph P. Danis, and John J. Carey to attorney records in our legal directory.

We use regular expression patterns to extract names and name matching evidence which includes law firm, city, and state information. Our name matching evidence consists of features that compare each of the following fields: first name, middle name, last name, firm name, and city/state. The values of the features are: matches exactly, matches in a fuzzy way, is unknown, or mismatches. An example of fuzzy matching would be if one name is a nickname of the other or if one name is an initial only and matches the first letter of the other name.

We use several thousand positive training examples to train a naïve Bayes match classifier. The size of our database was approximately 1 million records. We blocked on last name first, and, if we failed to find a match with this block, we blocked on first name. This multiple blocking method allowed us to capture cases where an attorney has changed her last name through marriage for example.

We compared our method to three other matching techniques for an attorney name. We measured the precision and recall we would get (1) if we link attorney names only when the first, middle, last name, and city-state match exactly, (2) if we link attorney names only when the first, middle, and last name match exactly without regard to city-state or firm information, and (3) if we link attorney names only when the first and last name match exactly without regard to middle name, city-state, or firm. The results are shown below and are compared with the naïve Bayes matching. As can be seen, the naïve Bayes technique significantly outperforms the baseline methods. For this comparison, we used a single match threshold of 0.25. High template and database record pairs scoring above the threshold were considered matched and those falling below were considered to signify an unmatchable template record.

	Prec.	**Recall**	**F**
Naïve bayes with threshold 0.25	0.993	0.916	0.953
Exact Match on first name, middle name, last name, and city-state	0.994	0.422	0.592
Exact Match on first, middle and last name	0.950	0.613	0.745
Exact Match on first and last name only	0.939	0.590	0.725

Table 1: Attorney matching methods comparisons

2.2 Case study: linking persons, companies, and locations from financial newswires to corresponding directory listings

We have also tagged mentions of companies, locations, and persons in financial news text and resolved them to corresponding authority files. Our biggest challenge in this application has been the resolution of persons. Our authority file consists of 677,765 person records: the officers and directors of publicly traded companies.

Our template record consists of the first, middle initial, last name, and companies named in the article. We block using the first and last name of the record. The blocks contain 4 or less records 96% of the time; however, some contain over 80 records. The matching phase is performed using a set of heuristics. Rules for positive resolution are applied in order of greatest-to-least evidence and confidence. Measures of evidence and confidence include the degree to which a name mention in the text is an exact match with the authority file and whether or not the company name associated with a particular name record is also mentioned in the document text. Names that are common with respect either to having many records associated with them, or in terms of a measure of overall name commonness (as determined by counts in a credit header database) are considered to be low-confidence and require more evidence for positive resolution.

Our system achieves an F-measure of 92.2% on person resolution (91.7% precision, 92.7% recall). This can be compared against a baseline of 50% accuracy. This baseline is produced by randomly choosing a match from the block which average 2 records in size.

3. Mapping records in one database to those in another database

We consider one of the databases to be the target and then, as in the previous section, the task of matching records in a database with those in the target database consists of the two phases mentioned in the previous section: blocking and matching.

Blocking can be explained in terms of extracting sets of candidate records from the target database that satisfy certain query parameters — the goal of which is to select only those blocks of data that meet certain requirements for further processing (e.g., last name matches query AND zip code matches query). When a given blocking function does not yield any candidate match, a broader blocking function is tried. Matching is done by scoring a feature vector of similarities over the various fields. The feature values can be either binary (verifying the equality of a particular field in the update and a master record) or continuous (some kind of normalized string edit distance between fields like *street address*, *first name,* etc).

As in the previous section, the evaluation of such a matching task typically includes precision and recall in an IR sense, as well as the associated F-measure. We may also wish to measure our progress in terms of precision among the non-matches (how often is our "don't match" decision correct)? Speed in terms of resolutions-per-second is another metric that real-time production applications often monitor.

3.1 Case study: the physician database

The task consists of merging a physician record from an *"update" database* to the record of the same physician in a *master record database*. The update database has fields that are absent in the master record database and *vice versa*. The fields in common include the *name* (first, last and middle initial), several *address* fields, phone, specialty, and the *year-of-graduation*.

More specifically, the system merges each of 20,000 physician records to the record of the same physician in the *master record database* consisting of approximately 1 million records. The fields in common include the *name* (first, last and middle initial), several *address* fields, phone, specialty, and the *year-of-graduation*.

Although the *last name* and *year of graduation* are consistent when present, the *address*, *specialty* and *phone* fields have several inconsistencies owing to different ways of writing the address, new addresses, different terms for the same specialty, missing fields, etc. However, the *name* and *year* alone are insufficient for disambiguation. We had access to ~500 manually matched update records for training and evaluation (about 40 of these update records were labeled as unmatchable with the information available).

We performed blocking by querying the master record database with the *last name* from the update record. Matching was done by scoring a feature vector of similarities over the various fields. The feature values were either binary (verifying the equality of a particular field in the update and a master record) or continuous (some kind of normalized string edit distance between fields like *street address*, *first name* etc.).

The logistic-regression-based matching algorithm assigns to each feature vector the probability that it corresponds to a match. All the records in the block are ranked according to this probability and the highest scoring record is assigned as the match if its score exceeded some appropriate threshold.

The training of the logistic regression algorithm was done by a semi-supervised algorithm called *surrogate learning*, which is based on the property that the binary *year of graduation* feature is independent of the other features if the two records are not matches. The reader is referred to (Veeramachaneni & Kondadadi, 2008) for a description of the algorithm and experimental results.

The matching algorithm was evaluated on 500 manually matched records with n-fold cross-validation. From this assessment, the precision and recall of the algorithm were determined to be 96% and 95% respectively.

4. Clustering records in a single database

In some cases, a single database table contains many records for the same entity but there is no explicit link expressing the identity relationship. The task then is to partition the table into equivalence classes where each class contains all the records for a specific entity. Again the task breaks down into the subtasks of blocking and matching; however, a third task of clustering is also required. We have successfully employed the similar blocking and matching techniques to those described in the previous sections. For clustering, we have used agglomerative clustering but other methods could also be employed (Jain & Dubes, 1988).

Evaluation, by contrast, does not follow the approach of the previous tasks. Instead of statistics based on counts of record pair linkages correctly found, incorrectly proposed, missed, etc., the statistics are based on counts with in clusters and then averaged over clusters.

4.1 Case study: account rolling

Within one of our internal accounting systems, multiple database records may exist for a single customer. Each record corresponds to a separate license for a single product. The customer database totals approximately 1.5 million records. The record format allows for flexibility in identifying the customer: up to four text fields may be used to name the customer entity, contact entity, and secondary entities such as departments, offices, regions, etc. The database is populated by multiple systems and consistent text field usage is not enforced. To help facilitate the assignment of sales representatives, the application needs to resolve account clusters by customer, using textual information only (the four name fields and address fields). Customer types include corporations, state and federal governmental agencies, and educational institutions. Corporate names tended to vary over time, reflecting mergers. Governmental customer names could also be non-unique: the same name may be utilized by similar entities in different cities, counties, states, and federal jurisdictions.

The database did indicate the market segment, if known, of the record. Therefore, clustering could be performed within each segment separately. Two thirds of the records had a non-null market segment. Unknown records were to be matched against the resulting segment clusters and added if matched.

The large corporations were expected to produce a relatively small number of large population clusters. A typical large corporation might have several hundred accounts. Approximately 50,000 accounts were expected to produce about 250 clusters. Far more problematic were the state governmental accounts. These represent the largest number of records, over 350,000. Clusters were expected to be numerous and very sparsely populated.

An SVM was used to compare record pairs. The feature data in each segment varied in completeness, location, and structure. In each of the segments, we wanted to match and cluster on the name of the entity. Feature selection involved selecting the optimum combination of the four text fields for each segment to determine the best cross match between records to keep expensive string comparisons to a minimum. The Jaro-Winkler algorithm was predominantly used in order to weight the first part of the string.

The SVM was trained on user provided gold data pairs. We selected a ratio of positive to negative training pairs of 1/2 (2000 and 4000 pairs respectively were used); 80% of the sampled pairs were used for training and the remaining 20% used for model validation. We performed validation experiments to select the optimal combination of SVM parameters (C, gamma, and kernel). An RBF kernel was used.

A basic agglomerative clustering technique was employed. The first record was set aside as the first cluster. The second record was compared to the first. If it matched (the SVM score exceeded a configurable threshold), it was added to the cluster. Otherwise a new cluster was created. Each subsequent record in the input data set was compared to existing clusters. When comparing a record to a cluster, the record was compared to each record in the cluster until either a match was found that exceeded the threshold or a negative match was found. If there were more than one matched cluster, the matched clusters were merged together.

After all of the records had been processed, the single valued clusters (i.e. clusters with only one element) were extracted and re-run through the process using the multi-valued clusters as the starting point. This was repeated until the number of single valued clusters reached equilibrium.

A final cluster merging was performed on the multi-valued clusters. The most frequently occurring entity name in each cluster was determined. For any two clusters, if the they had the same majority entity name and a similarity score (the product of the ratios of the number of occurrences of the majority entity name to the number of records in the cluster - a modified cosine similarity) between the two exceeded a configurable threshold (usually .80), the two clusters were merged.

Standard precision and recall metrics lacked a precise definition when applied to clustering. We initially devised two related metrics, purity and fragmentation to compare our cluster results with the gold data. Purity, a measure of how many records in the cluster belong together, measures the precision of the clusters at both the macro and micro level. Fragmentation attempted to quantify how many clusters it took to represent the true cluster. Purity is defined with respect to the generated clusters and fragmentation is defined with respect to the gold standard clusters. A purity of 1 and a fragmentation of 0 would indicate a perfect cluster.

The fragmentation scores were not informative enough. Similar fragmentation scores did not indicate how and to what extent the records were distributed across the set of clusters. A detailed tabular approach provided much better measurements:

Let **G** be a gold data cluster:
the set off all accounts, $\mathbf{a_i}$, that belong to a single customer.

Let **C** be the set of all generated clusters that completely enclose **G**:

for all $\mathbf{a_i}$ in **G**, $\mathbf{a_i}$ is a member of a cluster in **C**

Fragmentation of **G** equals the number of clusters in **C - 1**

Let $\mathbf{C_j}$ be a generated cluster:

Purity of $\mathbf{C_j}$ equals (size of largest gold standard contributor to the cluster) / (size of $\mathbf{C_j}$)

For any given sample, we determined the gold data clusters (record ids and count). For each gold data cluster, we found all generated clusters that contained an occurrence of a record id. For each of these clusters, we calculated the coverage ratio of id occupancies to the size of the cluster. For the three largest clusters, we reported the coverage ratios (this is a measure of how well any one of these clusters covers the target gold data cluster). We then accumulated average coverage scores for all clusters and macro coverage scores over the entire sample. We also reported the number of times a single cluster is generated that exactly covers the corresponding gold data cluster.

For each of the three largest clusters reported on for each gold data cluster, we calculated the purity of the cluster by taking the ratio of correct matches to the size of the cluster, then accumulated both micro and macro averages.

Let us now apply these metrics to our system's output. When compared against the customer's existing method of clustering (a rule based system), we produced higher coverage scores for the largest generated cluster. We placed more records in a single large cluster while the existing method tended to distribute records over two or more large clusters. Both approaches had residual single records. Purity scores were consistently high (0.99 for large sized clusters) so the comparison and clustering techniques were valid. Nonetheless, fragmentation could not be reduced due to insufficient evidence in the remaining single valued clusters.

Other comments on the output:

- There were a large number of single records that could not be clustered. In most cases, a valid entity name was missing (not present in any of the four possible record fields) or only a contact name (a person) was entered. The appearance of just a person name caused over-rolling (records placed in the wrong cluster) because of similarity of the person names (filtering techniques removed most of these problems).

- The entity names in governmental segments were not unique. The same name could indicate both a match and a mismatch. For example "Court Magistrate" was a match within the same circuit court, but a mismatch otherwise (this also resulted in positive and negative training vectors that were identical).

- There were a large number of single records that could not be merged into their respective clusters. This results in large fragmentation (e.g. we could generate one large cluster that covered 90% of the records in a gold data cluster, but the remaining records resulted in single clusters that could not be merged).

- Collections that are comprised of a relatively small number of large clusters are best suited to our techniques. Collections that consist of a very large number of very small or singular clusters did not perform as well. It looks like our techniques did quite well when the clusters were large enough to establish strong similarity measurements between records. For sparsely populated clusters, there wasn't enough evidence.

5. Summary

We have described a number of entity record tasks. The first two tasks were (i) linking mentions of people and companies in legal text to structured authority files and (ii) linking mentions of entities in financial newswires to structured authority files. The next task involved mapping records in one database to those in another: record matching for a physician database. Finally, we described the task of clustering record accounts so that clusters contained all accounts for a single company.

Although each of the entity record linking, mapping, and clustering problems described above are distinct, and invite their own innovative solutions, there also exists among them some common dimensions and broader lessons to be learned. Some of these common dimensions include the following. In an IR-like manner, there exists a clear trade-off between precision and recall. One generally cannot make dramatic gains in one without witnessing degradation in the other. It may only be the ratio of the benefit-to-cost that may change (e.g., a two point gain in recall costing five points in precision). Just as significantly, precision and recall only tell part of the story, and tend to understate other challenges associated with the problem space, for instance, deciding that a candidate pair does *not* represent a solid match (i.e., avoiding false positives, a.k.a., *non-match* precision) can be just as challenging as deciding that a match is validated. Other auxiliary metrics like average block size, in the case of linking or mapping, or maximum obtainable coverage or purity, in the case of clustering, can be equally informative indicators of problem difficulty or solution quality, and cannot be ignored when striving for globally optimal solutions. Still other issues carry additional lessons relating to the scale of the problem, the diversity of the available data sources, and the dynamic nature of the underlying entity data. Each of these dimensions compound the entity resolution challenge, and require real-world solutions in order to satisfy the underlying practical constraints. Because our solutions are focused on industrial applications, results that surpass existing baselines but ignore these critical dimensions (scale, varying record quality, dynamic environments) are not acceptable. Ultimately these approaches need to deliver high performance solutions in terms of result quality, scalability, and robustness, not to mention speed.

6. References

Baxter, R., Christen, P., and Churches, T. (2003). A Comparison of Fast Blocking Methods for Record Linkage. In *Proceedings of ACM SIGKDD 2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation.* Washington, D.C., USA.

Dozier, C. and Haschart, R. (2000). Automatic Extraction and Linking of Person Names in Legal Text. In *Proceedings of RIAO 2000 (Recherche d'Information Assistee par Ordinateur).* Paris, France: pp. 1305--1321.

Jain, A.K. and Dubes, R.C. (1988). Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, NJ.

Veeramachaneni, S. and Kondadadi, R.K. (2008). *Surrogate Learning—From Feature Independence to Semi-supervised Classification.* Submitted to ICML 2008.

Winkler, W. (1995). Matching and Record Linkage. In Cox, B. G., ed., *Business Survey Methods*, Wiley.

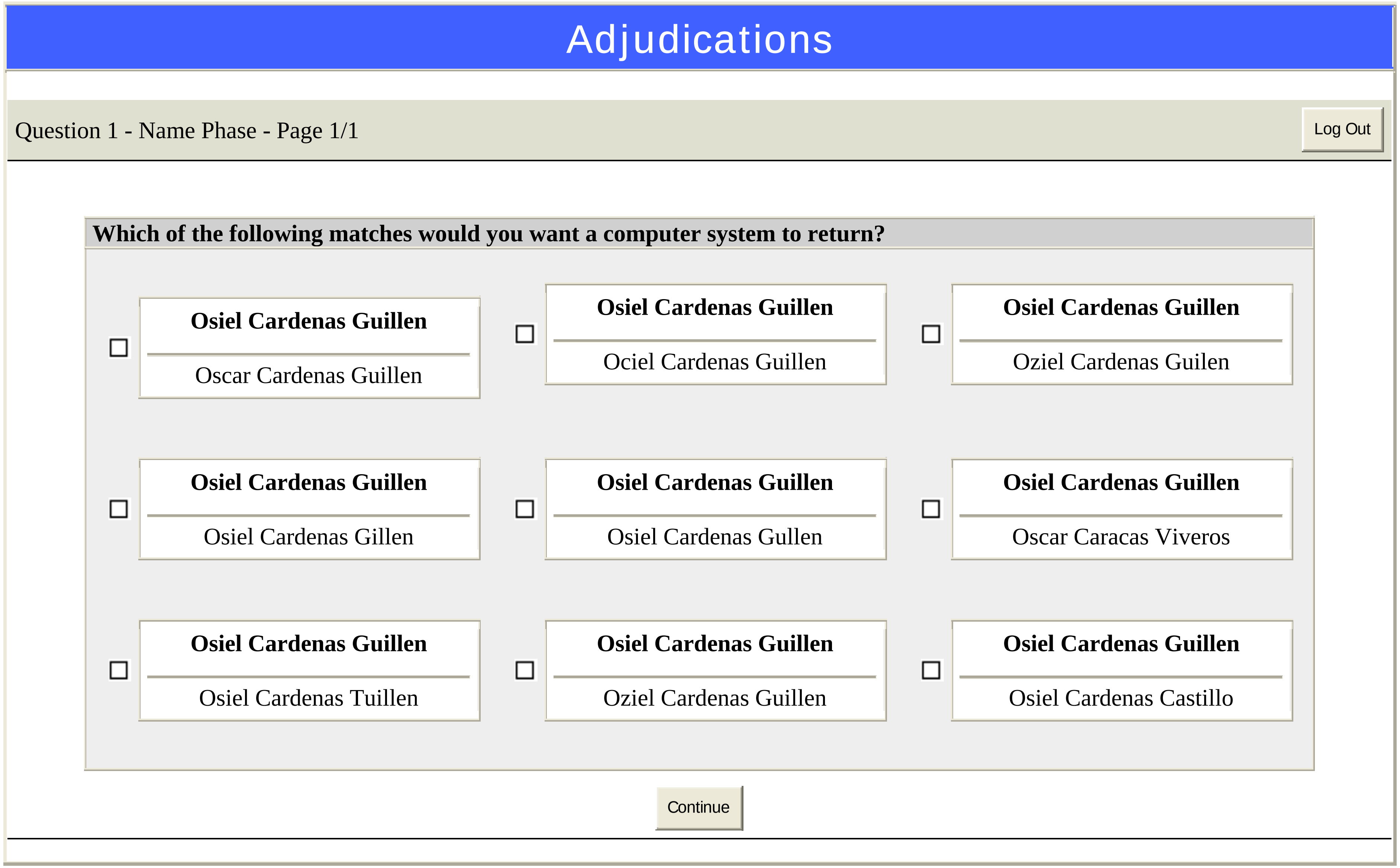
LREC Identity Resolution Workshop
Name Matching Exercise
May 31, 2008

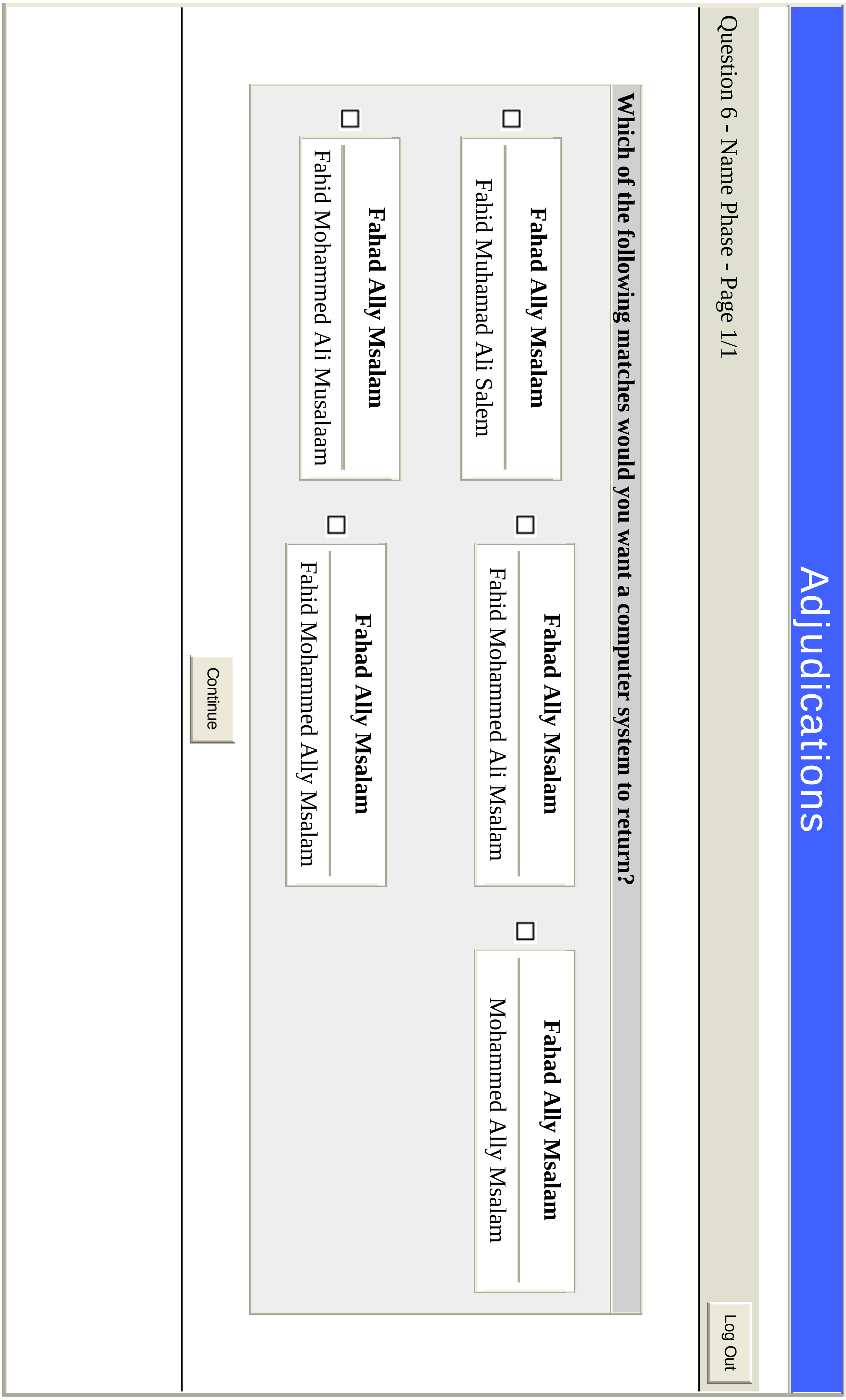
The names below were extracted a couple years ago from http://www.ustreas.gov/offices/enforcement/ofac/sdn/.

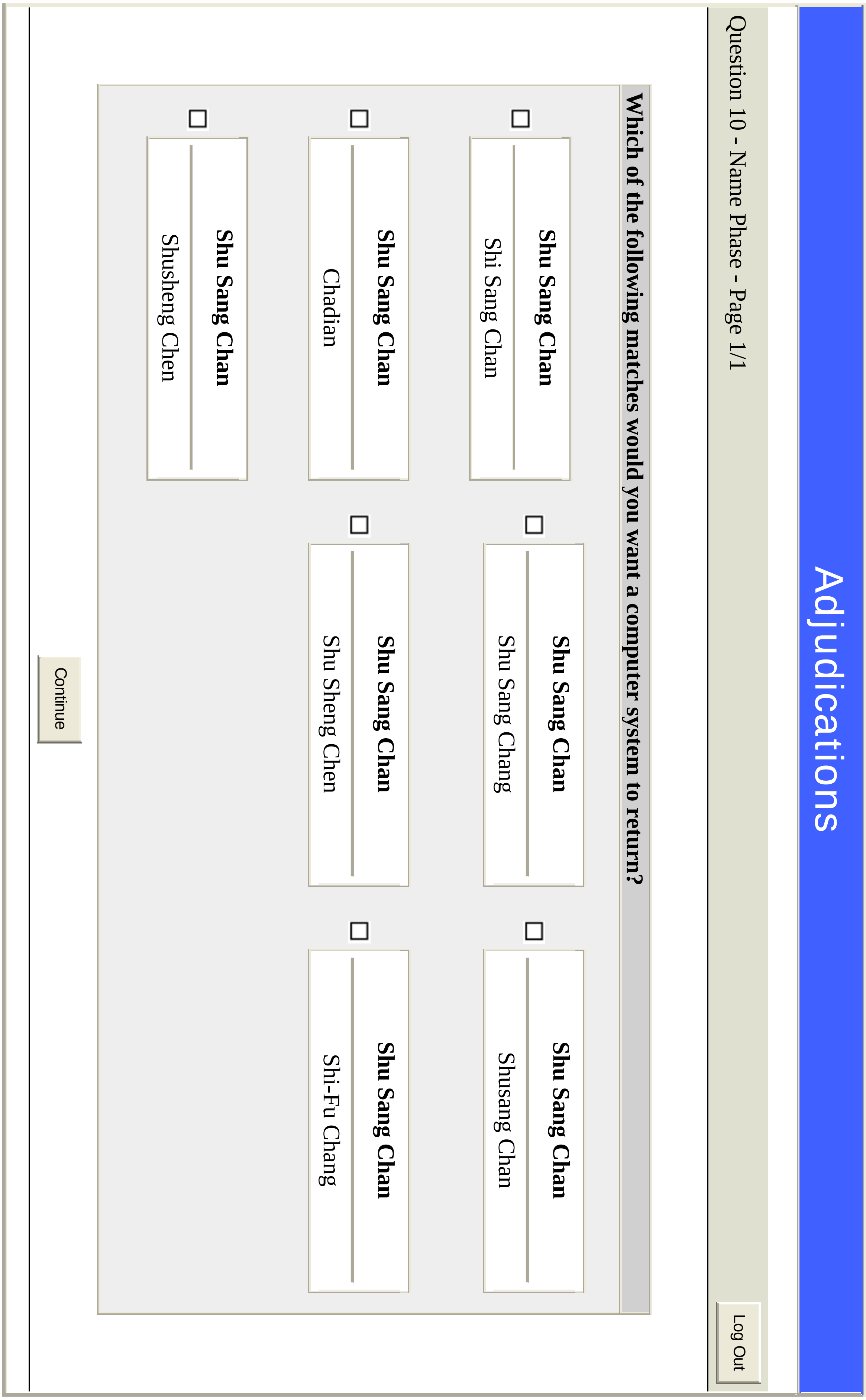
Which matches would you want a search engine to return?

Match?	No.	Query Name	Database Name
	1.	Sia-Kang Wei	Hsueh Kang Wei
	2.	Sia-Kang Wei	Shao-Kang Wei
	3.	Sia-Kang Wei	Xuekang Wei
	4.	Mahmoud Diab Al-Ahmad	Abu Ahmad
	5.	Mahmoud Diab Al-Ahmad	Ahmed the Tanzanian
	6.	Mahmoud Diab Al-Ahmad	Mahmud Dhiyab Al-Ahmad
	7.	Oscar Malarbe	Oscar Mahlerbe
	8.	Oscar Malarbe	Oscar Malherbe De Leon
	9.	Oscar Malarbe	Oscar Malmerbe
	10.	Oscar Malarbe	Oscar Macherbe
	11.	Oscar Malarbe	Oscar Malerva
	12.	Oscar Malarbe	Oscar Qalharbe De Leon
	13.	Oscar Malarbe	Oscar Ramirez M.
	14.	Oscar Malarbe	Oscar Nalherbe
	15.	Hadj Ahmed Nasreddin	Hajj Ahmed Salahaddin
	16.	Hadj Ahmed Nasreddin	Ahmed Idris Nasreddin
	17.	Hadj Ahmed Nasreddin	Ahmad I. Nasreddin
	18.	Barzan Ibrahim Hassan Al-Tikriti	Barzan Ibrahim Hassan Al-Takriti
	19.	Barzan Ibrahim Hassan Al-Tikriti	Ali Barzan Ibrahim Hasan Al-Tikriti
	20.	Barzan Ibrahim Hassan Al-Tikriti	Barzan Brahim Hassan Tikriti
	21.	Barzan Ibrahim Hassan Al-Tikriti	Mohammad Barzan Ibrahim Hasan Al-Tikriti
	22.	Barzan Ibrahim Hassan Al-Tikriti	Sabawi Ibrahim Hassan Al-Takriti

Match?	No.	Query Name	Database Name
	23.	Nasir Ali Khan	Nazir Ali Khan
	24.	Nasir Ali Khan	Nasran Khan
	25.	Nasir Ali Khan	Nafir Ali Khan
	26.	Nasir Ali Khan	Ali Khan
	27.	Nasir Ali Khan	Nisar Ali Khan
	28.	Nasir Ali Khan	Nisan Ali Khan
	29.	Nasir Ali Khan	Naser Alfred Khant
	30.	Winai Pichayos	Vinai Pitchayos
	31.	Winai Pichayos	Vinai Tichyos
	32.	Winai Pichayos	Vinai Pichayot
	33.	Winai Pichayos	Winai Phitchaiyot
	34.	Winai Pichayos	Winai Thichaiyot
	35.	Dhu Himma Shaleesh	Zuhilma Shalish
	36.	Dhu Himma Shaleesh	Dhu Himma Saleeb
	37.	Dhu Himma Shaleesh	Dhu Al Himma Shalish
	38.	Dhu Himma Shaleesh	Dhuil Himma Shalish
	39.	Dhu Himma Shaleesh	Thu Al Hima Shaleesh


Adjudications
Question 1 - Name Phase - Page 1/1
Log Out
Which of the following matches would you want a computer system to return?
Osiel Cardenas Guillen
Oscar Cardenas Guillen
Osiel Cardenas Guillen
Ociel Cardenas Guillen
Osiel Cardenas Guillen
Oziel Cardenas Guilen
Osiel Cardenas Guillen
Osiel Cardenas Gillen
Osiel Cardenas Guillen
Osiel Cardenas Gullen
Osiel Cardenas Guillen
Oscar Caracas Viveros
Osiel Cardenas Guillen
Osiel Cardenas Tuillen
Osiel Cardenas Guillen
Oziel Cardenas Guillen
Osiel Cardenas Guillen
Osiel Cardenas Castillo
Continue


Adjudications
Log Out
Question 6 - Name Phase - Page 1/1
Which of the following matches would you want a computer system to return?
Fahad Ally Msalam
Fahid Muhamad Ali Salem
Fahad Ally Msalam
Fahid Mohammed Ali Msalam
Fahad Ally Msalam
Mohammed Ally Msalam
Fahad Ally Msalam
Fahid Mohammed Ali Musalaam
Fahad Ally Msalam
Fahid Mohammed Ally Msalam
Continue


Adjudications
Log Out
Question 10 - Name Phase - Page 1/1
Which of the following matches would you want a computer system to return?
Shu Sang Chan
Shi Sang Chan
Shu Sang Chan
Shu Sang Chang
Shu Sang Chan
Shusang Chan
Shu Sang Chan
Chadian
Shu Sang Chan
Shu Sheng Chen
Shu Sang Chan
Shi-Fu Chang
Shu Sang Chan
Shusheng Chen
Continue

LREC Identity Resolution Workshop
Entity Resolution Confusion Corpus
May 31, 2008

The following web documents reflect the kinds of pages that can be found for three names.

How many entities are named Martin Jones in the documents?

How many entities are named Michael Taylor?

How many entities are named Sharon Smith?

March 2, 2008 30th Annual Napa Valley Marathon

The Course :: Course Records

Division 19 and Under

Men			
Mike Warr	18	2:31:21	1980
Michael Dudley	19	2:31:21	1990
Tim Lee	19	2:48:14	1979
Ernest Price	18	2:49:10	1981
Timothy Grove	18	2:54:23	2000
Women			
Kristie Clemens	19	3:13:10	1989
Mandi Reynolds	19	3:13:34	1997
Kathy D'Onofrio	18	3:14:05	1983
Anne Hitchcock	19	3:20:42	1998
Emilee Del Valle	17	3:25:11	1998

Division 20 - 24

Men			
Jamie White	23	2:16:34	1980
Mike Warr	21	2:22:52	1983
Chris Ashfield	23	2:24:03	2000
Dean Rinde	23	2:24:19	1987
David Chairez	24	2:24:29	1984
Women			
Eileen Kraemer	24	2:53:30	1984
Kathleen Smith	21	2:54:33	1988
Cristy Runde	24	2:56:51	1993
Megan Daly	21	2:58:17	2000
Hillary Simmons	20	2:59:36	1990

Division 25 - 29

Men			
Brent Friesth	27	2:18:28	1988
David Chairez	27	2:18:58	1988
Joseph Karnes	28	2:21:08	1994
Dean Rinde	26	2:24:07	1990
Doug McLean	27	2:24:54	1981
Women			
Betsy Swan	26	2:46:41	1991
Joanne Ernst	25	2:47:05	1984
Jeannie Urness	29	2:47:17	1992
Ann Trason	27	2:47:20	1988
Mariam Schmidt	29	2:47:24	1999

Division 30 - 34

Men			
Dick Beardsley	30	2:16:20	1987

Thomas Borschel	30	2:21:04	1988
Dan Aldridge	33	2:21:42	1990
Craig Morre	33	2:21:54	1987
Aaron Pierson	32	2:23:58	1996

Women

Diana Fitzpatrick	33	2:39:42	1992
Chris Iwahashi	34	2:46:49	1990
Peggy Smyth	30	2:51:01	1983
Sharlet Gilbert	31	2:51:50	1982
Cheryl Boessow	34	2:51:54	1995

Division 35 - 39

Men

Charles Thompson	35	2:25:50	1985
Eoin Fahy	37	2:25:53	1997
Eoin Fahy	38	2:28:53	1998
Paul Bonfiglio	35	2:29:05	2000
Chris Clark	37	2:29:44	1997

Women

Ann Trason	38	2:45:39	1999
Ann Danzer	36	2:47:30	1984
Wendy O'Donnell	38	2:51:00	1982
Chris Iwahashi	35	2:53:05	1991
Nellie Wright	37	2:54:04	1983

Division 40 - 44

Men

Richard Flores	44	2:25:52	1999
Richard Flores	41	2:26:04	1996
Rob Reid	41	2:27:40	1996
Jeffrey Wall	41	2:30:39	1994
Gustavo Figueroa	42	2:30:56	1994

Women

Marilyn Harbin	43	2:54:46	1981
Joan Ullyot	43	2:55:20	1984
Joan Reiss	44	2:57:24	1982
Elizabeth Sonne	41	2:58:51	1988
Diane McEven	40	2:58:33	1983

Division 45 - 49

Men

Ken Wilson	45	2:31:38	2000
Charles Thompson	45	2:32:38	1995
Martin Jones	45	2:37:49	1990
Darryl Beardall	46	2:39:13	1984
Will Pittenger	46	2:45:18	1997

Women

Joan Ullyot	48	3:07:32	1989
Susan Kielsmeier	46	3:13:40	2000
Philomena Chandra	45	3:16:50	1998
Corky Keefe	46	3:16:56	1989

Dick Yeager	66	3:38:34	
Women			
Myra Rhodes	65	3:44:24	
Myra Rhodes	69	3:57:30	
Peggy Hansen	67	4:30:56	
Marlene Kinser	65	5:30:00	
Peggy Ewing	68	5:36:28	

Division 70 - 74

Men			
Don Lundberg	73	3:35:57	
Paul Reese	71	3:41:49	
Max Jones	71	3:42:04	
Harrie Hess	71	3:58:15	
G. Billingsley	71	4:00:13	
Women			
Marci Trent	70	4:11:54	
Helen Klein	70	4:23:51	
Marvis Lindgren	73	4:34:08	
Helen Klein	72	4:46:53	
Etta Palmer	70	5:09:50	

Division 75 - 79

Men			
John Keston	77	3:34:48	2002
John Keston	78	3:36:41	2003
G. Billingsley	75	4:39:33	
Charles Hoagland	75	5:20:26	
Charles Hoover	76	5:30:00	
Women			
Helen Klein	79	4:48:06	

Division 80 +

Women		
Helen Klein	80	4:41:53

SUMMIT BAPTIST ASSOCIATION
Wednesday, April 09, 2008

- Skip Navigation
- Home
- SBA Congregations
- SBA Meetings
- SBA Updates
- SBA Projects
- Meet Martin Jones
- Need a Preacher, Speaker, or Music?
- Ministry E-Sources
- SBA Team Leaders
- Men's Events
- Women's Events
- Events
- Contact Us

Make this my home page.

Meet Martin Jones

Bible Search

KJV GO

Add Bible to your site

Martin Jones is the Summit Baptist Associational Missionary

Martin, his wife Karen, and their family reside in Canal Fulton. Evangelism has always been a major part of any strategy of growth in Martin's ministries. One of his first opportunities to assist churches in evangelism occurred while completing his seminary education. After training in Continuing Witness Training, he began and led several sessions of CWT at Riverside Baptist Church, Fort Worth, Texas, as Pastor/Leader.

Personally sharing his faith and leading others to share theirs continued in his ministries as he started two Church Plants. His first church plant, Northside Baptist Church in Huntsville, Texas, grew from a home Bible study to a church of 50 committed members. He started Eastview Baptist Church in Mesquite, Texas, with 14 people, and in two years the church had an average attendance of 60 and a new church building.

While working as a chaplain for the Metropolitan Detention Center in Los Angeles, CA, Martin was responsible for the preservation of inmate First Amendment Rights. This experience gave him an opportunity to work with various religious groups and to discover methods of sharing his faith in non-threatening ways.

When he became pastor of Brea Center Church in Brea, California, the church had a median age of 63 in a community with a median age of 34. Martin needed to reach people for God and he needed to reach them fast. He led the church to develop an evangelist strategy called Vision 2000 and Beyond and as a result the average worship attendance grew by 70%, small group/Sunday School ministry grew 150%, and giving went up 48%. Also during this time, the median age of the church went from 63 to 34 years of age.

Powered by E-zekiel v.2.6

Your browser does not support script

Department of Psychiatry

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Home | Resources for Employees | Contacts | Search

Services:
Clinical

Programs:
Educational
Research

Faculty Directory

Michael J. Taylor, Ph.D.
Assistant Professor of Psychiatry
E-mail: mjtaylor@ucsd.edu
PHONE #: (858) 642-3101
FAX #: (858) 552-7432

Biography

A long-term resident of San Diego, Dr. Taylor received his B.A. in psychology from UCSD in 1989 and his M.A. in psychology from SDSU in 1991. After his internship at the VA Connecticut Healthcare System /Yale Clinical Campus, he earned a PhD in clinical psychology with a specialization in neuropsychology from the SDSU/UCSD Joint Doctoral Program in Clinical Psychology in 1996. Dr. Taylor completed a postdoctoral internship at UCSD and is currently an Assistant Adjunct Professor in the Department of Psychiatry at UCSD and a member of the SDSU/UCSD Joint Doctoral Program in Clinical Psychology faculty.

Research Focus

Dr. Taylor's primary research goal is to apply magnetic resonance spectroscopy (MRS) and other novel neuroimaging techniques to the study of diseases impacting the CNS in order to evaluate treatment efficacy and/or disease progression. He is currently conducting three NIMH funded studies tracking the brain changes associated with HIV treatment. He is also the lead investigator of a VA funded study of the CNS consequences of alcoholism measured with MRS, diffusion tensor imaging, and cognitive testing.

Clinical Focus

Dr. Taylor is a licensed clinical psychologist, with specific interests in the generation and application of demographically-corrrected norms in neuropsychological assessment. He is also a member of the Disaster Mental Health Services team for the San Diego Chapter of the American Red Cross.

Selected Publications

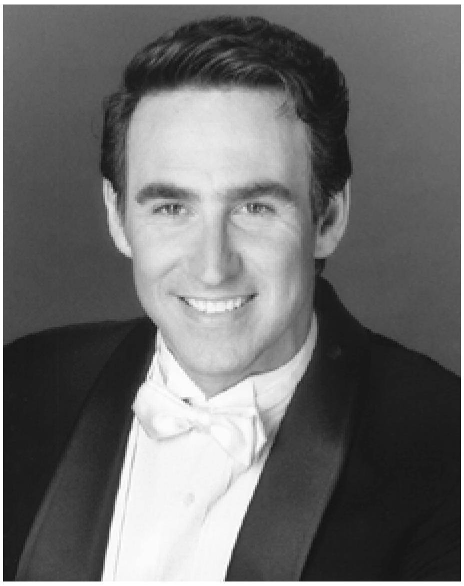
- M. J. Taylor, O. M. Alhassoon, B. C. Schweinsburg, J. S. Videen, I. Grant, & the HNRC Group. "MR Spectroscopy in HIV and Stimulant Dependence." Journal of the International Neuropsychological Society, 6, 2000 (pp. 83-85)
- B. C. Schweinsburg, M. J. Taylor, O. M. Alhassoon, J. S. Videen, G. G. Brown, T. L. Patterson, F. Berger, & I. Grant. "Chemical Pathology in Brain White Matter of Recently Detoxified Alcoholics: A 1H Magnetic Resonance Spectroscopy Investigation of Alcohol-Associated Frontal Lobe Injury." Alcoholism: Clinical and Experimental Research, 25, 2001 (pp. 924-934)
- M. J. Taylor, & R. K. Heaton. "Sensitivity and Specificity of WAIS-III/WMS-III Demographically Corrected Factor Scores in Neuropsychological Assessment."

Journal of the International Neuropsychological Society, 7, 2001 (pp. 867-874)

- M. J. Taylor, S. L. Letendre, B. C. Schweinsburg, O. M. Alhassoon, G. G. Brown, A. Gongvatana, I. Grant, I., & the HNRC Group. “Hepatitis C virus infection is associated with reduced white matter N-acetylasparate in abstinent methamphetamine users.” Journal of the International Neuropsychological Society, 10, 2004 (pp. 110-113)
- B. C. Schweinsburg, M. J. Taylor, O. M. Alhassoon, R. Gonzalez, G. G. Brown, R. J. Ellis, S. Letendre, J. S. Videen, J. A. McCutchan, T. L. Patterson, I. Grant, & the HNRC Group. “Brain mitochondrial injury in human immunodeficiency virus-seropositive (HIV+) individuals taking nucleoside reverse transcriptase inhibitors.” Journal of Neurovirology, 11, 2005 (pp. 356-364)

University of California, San Diego, Department of Psychiatry, 9500 Gilman Drive, Mail Code 0603 La Jolla, CA 92037-0603
Telephone: (858) 534-3684, Fax: (858) 534-7653, Electronic Mail: psychiatry@ucsd.edu

Michael Taylor



Michael Taylor sings the role of Stefano in *Viva la mamma*. Mr. Taylor graduated from the San Francisco Conservatory of Music with his master's degree in 1990. He has appeared as soloist with many opera companies including San Francisco Opera, Sacramento Opera, Opera San José, Marin Opera, and West Bay Opera, singing such roles as Gianni Schicchi, Escamillo, the Count (*The Marriage of Figaro*), Scarpia, Dr. Malatesta, Don Giovanni, Belcore, Tonio, Figaro (*The Barber of Seville*), and many others. Mr. Taylor has appeared in concert with the Masterworks Chorale, Berkeley Symphony, Fremont Symphony, Sacramento Choral Society, and Schola Cantorum, and has performed as a vocal soloist with the San Francisco Ballet. A regional finalist in both the San Francisco Opera Merola Auditions and the Metropolitan Opera Auditions, Mr. Taylor was also a participant in the San Diego Opera Apprentice Program. Winner of the Bel Canto Foundation competition, Mr. Taylor spent six weeks in Siena, Italy, studying with coaches from La Scala. Mr. Taylor was also a member of the cast of Andrew Lloyd Webber's *Phantom of the Opera* at the Curran Theater in San Francisco.

Previous roles with West Bay Opera include the Count (*The Marriage of Figaro*), Valentin (*Faust*), Belcore *(Elixir of Love*), Germont (*La traviata* 1987), Marcello (*La bohème* 1986 & 1982), Scarpia (*Tosca* 1989 & 1984), Falke (*Die Fledermaus*), Tonio (*Pagliacci* 1985), and Silvio (*Pagliacci* 1978).

Updated September 16, 2003 by Lucinda Surber.

26 Results matching "Michael Taylor, San Diego, CA"

Display: All (26) | Home (21) | Work (5) **Sort by:** - Select -

1 2 3 Next >

Michael Taylor
2025 K St
San Diego, CA 92102-3853
(619) 238-4194
Listing Details

Michael A Taylor
3836 Alabama St, Apt 305
San Diego, CA 92104-3363
(619) 255-2760
Listing Details

Michael Taylor
4183 Mississippi St
San Diego, CA 92104-1629
(619) 260-1910
Listing Details

Michael N & Annie Taylor
3242 Wheat St
San Diego, CA 92117-4430
(858) 270-6341
Listing Details

Ages: 55-59, *unavailable*

Michael A & Susan G Taylor
7608 Topaz Lake Ave
San Diego, CA 92119-3046
(619) 460-3970
Listing Details

Ages: 45-49, 45-49

Michael Jr Taylor
8740 Donaker St
San Diego, CA 92129-4205
(858) 484-5488
Listing Details

Age: 30-34

Michael Taylor
4667 Torrey Cir
San Diego, CA 92130-6642
(858) 509-4731
Listing Details

Michael Taylor
1240 India St, Unit 1505
San Diego, CA 92101-8553
(858) 531-5180
Listing Details

Job title: Founder
Company: Eq8 Technologies

Michael P & Heidi A Taylor
13731 Via Tres Vis
San Diego, CA 92129-2732
(858) 538-0556
Listing Details

Ages: 45-49, 45-49

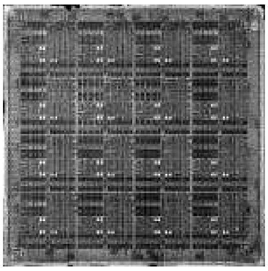
Michael E Taylor
11634 Timsford Rd
San Diego, CA 92131-3626
(858) 586-6333
Listing Details

Age: 60-64

Michael Taylor's Profile

REAL NAME™

Need help?
More information on Profile pages



Location: San Diego, CA

Customer Images: 1
See image gallery

Nickname:
mbt224

In my own words
I am not a compulsive liar.

Amazon Friends & Interesting People

Recently Active

Rose Kolodny added The Calligrapher's Bible: 100 Complete Alphabets and How to Draw Them by David Harris to Wish List.

Friends (3)


Bruce Jones


Rose Kolodny


Scott Silver

See Profile for: Bruce Jones GO

› See all 3 Amazon Friend(s)

Interesting People (1)


joy drake silver

Your Actions

- Invite as Amazon Friend
- Add to Interesting People
- E-mail this page

More to Explore

› Images (1) | Top Reviewers

Have feedback or suggestions about Amazon.com's community features?

About Sharon

Wednesday, April 09

Home
Classes
Workshops & Seminars
Calendar
Products
Comments & Testimonials
Retreats
Chi Nei Tsang
Taoist Inspiration
Chinese Astrology
Special Events
Mantak Chia
About Sharon

Sharon practicing at the Temple of Heaven in Beijing, 2007

SHARON SMITH has been practicing Qigong, Tai Chi, & other Taoist spiritual and healing arts for 29 years & teaching them for 24. Her influential teachers include Masters Mantak Chia (she was in his original class of western students in 1981), Li Jun Feng, and T.K. Shih. She also studied with Jeanette Chi, Gilles Marin, Don Ahn, as well as many other Tao masters. Sharon is certified by Master Mantak Chia's Universal Tao system as a Senior Instructor and Chi Nei Tsang Practitioner. She is also certified by the International Sheng Zhen Society to teach Sheng Zhen Wuji Yuan Gong, the work of Master Li Jun Feng. In addition, Sharon has practiced Iyengar yoga for over 20 years. She has traveled many times to China, Thailand, India, the Philpines, and New Zealand to further her studies.

Sharon currently teaches seminars internationally as well as regularly at the New York Open Center. She has also taught at The Tao Garden, Omega Institute, Healing Tao University, The Learning Annex, Wainwright House, United Nations Feng Shui Club, Madison Avenue Presbyterian Church, Morningside Retirement Health Services, Jubilee Senior Center, Healing Tao of New York, La Guardia Community College, Queens College, New York University, Adult Education Division of the New York City Board of Education, & the New York State Department of Parks & Recreation.

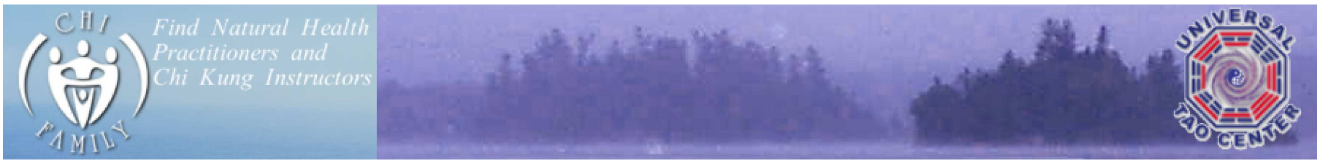
Sharon is a recipient of numerous foundation grants for her work with senior citizens and in community mental health programs. She was a spokesperson for Qigong on the nationally syndicated PBS television show, "Asian America". Sharon writes the Chinese Astrology Column for the Asian Food and Lifestyle Journal.

The Universal Tao

Founded by Tao Master Mantak Chia, the Universal Tao is the systematic study and practice of the the Natural Way (Tao) of healing and enlightenment. Master Chia has transmitted the hermit One Cloud's secret Seven Formulas of Immortality to the West and in the process, been instrumental in providing an opportunity for practitioners to synthesize the Taoist tradition with the latest scientific discoveries. The practices combine qigong (or chi kung) with the profound process and application of inner alchemy to enhance all areas of our life. In a variety of sitting, standing, and lying down practices we open the door to experience profoundly the timeless, practical wisdom of "The Way".

Sheng Zhen Wuji Yuan Gong

Sheng Zhen means "sacred truth" and refers to Unconditional Love Qigong which has been transmitted by the famous coach of the Beijing Wushi team, Master Li Jun Feng. This is a spiritual qigong composed of different sets of sitting and standing elegant movements and meditations inspired by the world's great spiritual traditions, This form of qigong has 3 functions -- to improve the body's health, to remove negative emotions and thoughts, and to open one's heart.



Home ▾ Read and Write ▾ Learn and Locate ▾ Look and See ▾ Play ▾ Contact ▾

Search

in:

Entire Site go

Meditations

Now Available: Sarina Stone's Guided Meditations for Download or on CD

Including

The Inner Smile
World Link Meditation
The Tao of Manifestation
The Tao of Cleansing
The Tao of Beauty
Free samples

Last articles

1) Domenic Thomas
2) Willow Lune
3) What Toxic Soup is Hidden in Your Bathroom?

Top Sites

1) Chi Kung: Energy Work
2) Baby Yoga
3) Essential Oils Breech
4) Sugar and Children
5) Quantum Dating Club

Featured links

Universal Healing Tao - Mantak Chia
Universal Healing Tao - Sarina Stone
Midwest Universal Healing Tao Practitioners and Educators
Directory by Topic

Last actions

Add Your Listing

Practitioners and Educators

We invite you to list your business here.

Contact us

Site Stats

Visitors to Date

13,111

Login

user:

pass:

login

Remember me

[register | I forgot my password]

Sharon Smith

By: System Administrator on: Aug 12, 2007 [13:42] (651 reads)

(2882 bytes)

About Sharon Smith:
SHARON SMITH has been practicing Qigong, Tai Chi, & other Taoist arts for 23 years & teaching them for 18. Sharon is a certified senior instructor of the Universal Healing Tao System, the work of Master Mantak Chia.

Sharon currently teaches at the New York Open Center as well as the Healing Tao of New York & has also taught at The Tao Garden, Omega Institute, Healing Tao University, The Learning Annex, Wainwright House, Jubilee Senior Center, Queens College, New York University, Adult Education Division of the New York City Board of Education, & the New York State Department of Parks & Recreation.

She was a spokesperson for Qigong on the nationally syndicated PBS television show, "Asian America".

Sharon Smith Born: Dallas, Texas, 1951 B.A., Philosophy, Vanderbilt University, Nashville, Tennessee Sharon Smith has been practicing Qigong, Tai Chi, & other Taoist arts for 24 years & teaching them for 19.

Her teachers include Mantak Chia, Li Jun Feng, Don Ahn, T.K.Shih, Yang Jwing Ming, Liang Shou Yu, Faxiang Hou, Lao Kang Wen, Ken Cohen, Fong Ha, Jeanette Chi & Juan Li. Sharon has also studied Hatha Yoga in the Iyengar method for 18 years with such noted teachers as Geeta Iyengar, Mary Dunn, Faiq Biria, Manouso Manos, & Ramanand Patel.

She has been to India & Thailand 4 times to further her studies. Sharon is a certified senior instructor of the Universal Healing Tao System, the work of Master Mantak Chia. Currently she teaches at the Healing Tao of New York & has also been a regular teacher at the New York Open Center.

Other teaching venues include Omega Institute, the Learning Annex, Wainwright House, Jubilee Senior Center, Morningside Reitrement & Health Services, Sanctuary for Families, Healing Tao University, Healing Tao Institute, Queens College, New York University, Adult Education Division of the New York City Board of Education, & the New York State Department of Parks & Recreation.

Sharon also teaches internationally & was a spokesperson for Qigong on "AsianAmerica", a syndicated PBS television show. References: Sandy Levine; Program Director; The New York Open Center; 212-219-2927, #132 Ronald Bruno; Executive Director; MRHS; 212-666-4000

How I Serve Children and Families:
I am interested in assisting the development of strong energetic connections between parents & children. Workshops & seminars are available for individuals as well as groups & organizations.

Sharon Smith
752 Greenwich St. #3C, NY, NY 10014 USA
Phone: 212-243-6771 Fax: 212-243-6771
www.taosharon.com
Taosharon@aol.com

home / subscribe / donate / books / archives / search / links / feedback / events / faq

"America's Best Political Newsletter" Out of Bounds Magazine

counterpunch

edited by alexander cockburn and jeffrey st.clair

March 26, 2008

Welfare on Wall Street

Greed Pays

By SHARON SMITH

On March 19, JPMorgan Chase chief executive Jamie Dimon joined Bear Stearns chief executive Alan Schwartz to face a group of 400 stunned Bear executives. Five days earlier, Bear Stearns, one of Wall Street's five largest investment banks, had lost $17 billion of wealth, triggering the biggest financial panic since the Great Depression.

Bear approached complete collapse before the U.S. Federal Reserve stepped in to rescue it by engineering the emergency funding that allowed commercial giant JPMorgan to take over Bear, the first time the Fed has engineered such a rescue since the 1930s.

Dimon and Schwartz somberly explained to the assembled executives, "we here are a collective victim of violence," as if the investment firm had been beaten and robbed by a gang of creditors instead of aiding and abetting its own rapid demise.

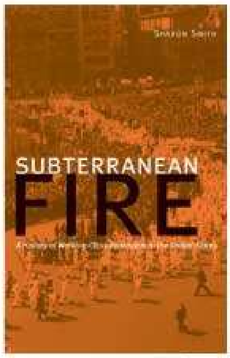
It is impossible to feel sympathy for the situation now facing Bear's high-flying management team. Schwartz continued to issue public assurances of Bear's solvency until the day the firm collapsed. Current non-executive chairman and former CEO Jimmy Cayne, who achieved billionaire status a year ago, has spent the better part of the last year attending to his hobby of card playing and was indeed at a bridge tournament in Detroit while the value of Bear stocks was evaporating last week.

Even now, Cayne will walk away with more than $16 million while JPMorgan has already reportedly made lucrative offers to hire top Bear bankers and brokers. Under pressure from Bear's board of directors, Morgan sweetened the pot, raising its initial offer of $2 per share to $10 on March 24-again winning praise from Schwartz.

Bear's 14,000 employees, in contrast, have fared poorly. They own an estimated one-third of its total shares, which only last year peaked at $171.50 per share. As Bear sheds half of its workforce, many will face financial ruin. The cost to workers whose pension funds have been invested in Bear Stearns is unknown.

"Wall Street is really predicated on greed"

The Bear Stearns debacle is just the latest phase of the financial distress triggered by the subprime mortgage crisis last July, and it is unlikely to be the last. In a moment of candor, former Bear board member Stephen Raphael summarized the unfolding crisis facing the U.S. financial system, telling the *Wall Street Journal,* "Wall Street is really predicated on greed. This could happen to any firm."



Subterranean Fire by Sharon Smith

The current financial panic is based on the knowledge that since the 1990s, Wall Street investment firms have orchestrated get-rich-quick schemes predicated on a model of betting using the odds of Russian Roulette, in which managers offer investors opportunities to make fast money in high risk transactions-through hedge funds, structured Investment Vehicles (SIVs) and other "innovative" derivative instruments such as Collateralized Debt Obligations (CDOs).

These investment schemes, which operate free of government regulation or oversight, have been described as a "shadow banking system," which operates in virtual secrecy, accountable to no one, based on mathematical models investors could not possibly understand and leveraged by borrowed money many times the actual money invested-at terms always skewed in favor of the short-term gains for managers.

The wheels for the current financial perfect storm were set in motion many years before the subprime mortgage crisis hit, and the Bush administration deserves no credit.

As one of his last acts as president in December 2000 Bill Clinton signed into law the Commodity Futures Modernization Act, which formally deregulated companies sponsoring derivatives schemes, sponsored by Texas Republican Phil Gramm, now the vice chairman of the Swiss investment firm UBS.

As *Financial Times* columnist Martin Wolf noted, "With the 'right' fee structure mediocre investment managers may become rich as they ensure that their investors cease to remain so."

On March 13, the Carlyle Capital Corporation hedge fund collapsed with debts amounting to 32 times its capital. The significance of Carlyle's demise was overshadowed by the Bear Stearns debacle. Yet, as Wolf argued, such vehicles are "bound to attract the unscrupulous and unskilled, just as such people are attracted to dealing in used cars

"It is in the interests of insiders to game the system by exploiting the returns from high probability events. This means that businesses will suddenly blow up when the low probability disaster occurs, as happened spectacularly at [the U.K. bank] Northern Rock and Bear Stearns."

Two of Bear Stearns hedge funds went under in the last six months due to disintegrating subprime mortgage holdings. But as the recent string of Wall Street crises exposed, the shadow banking system has increasingly intersected with commercial banks. It is difficult to know where one ends and the other begins, since banks have been allowed to keep such investment

Broken

Dave Zirin
Olympic Flames: China's Crackdown in Tibet

David Krieger
The Crisis in Tibet

Website of the Day
Memorializing Iraq

March 24, 2008

Jeffrey St. Clair
Blonde Ambition: Hillary's Berserker Campaign for 2012

Peter Morici
Digging Out of the Recession

Uri Avnery
Two Americas

Wajahat Ali
First of the Mohicans: an Interview with Rep. Keith Ellison

Paul Craig Roberts
Inside the Shell Game

George Ciccariello-Maher
The Coming War on Venezuela

Stephen Lendman
Sami Al-Arian's Long Ordeal

Christopher Brauchli
Possessing Someone Else's Country

Cat Woods
A Letter to Mom on Obama

Stacey Warde
Tax Burden

Dave Lindorff
The American Dead Hits 4,000, But Who's Counting?

Website of the Day
Live from the Longest Walk

March 22 / 23, 2008

Ralph Nader
Bush Blisters the Truth on Iraq

Nicole Colson

Can You Afford to Feed Your Family?

James Petras
The Cost of Unilateral Humanitarian Initiatives

Laura Carlsen
From Bombs to Markets: The Andean Crisis and the Geopolitics of Trade

Greg Moses
Tolerance and the American Pulpit

Andy Worthington
Torture Stories Dog Guantánamo Trials

Michael Dickinson
Art on Trial

John Ross
Bush's Surge Hits Mosul

Missy Comley Beattie
Killer Economics

David Michael Green
Happy Anniversary, America!

Ramzy Baroud
The Coming Uncertain War on Iran

Martha Rosenberg
Easter Egg Shells from Hell

Paul Watson
Evolution is Going to the Dogs in the Galapagos

Isabella Kenfield
Monsanto's Raid on Brazil

James Murren
Logging v. Water in Honduras

Jacob Hornberger
Sex and the Immigration Officer

Kathlyn Stone
Ben Heine, Master of the Art of Resistance

Seth Sandronsky
Rethinking New Mexico's History

Kim Nicolini
Class, Gender and Abortion in Communist Romania

Jeffrey St. Clair
Booked Up: What I'm Reading This Week

vehicles off their balance sheets--legally.

As the *New York Times* reported on March 23, "derivatives are buried in the accounts of just about every Wall Street firm, as well as major commercial banks like Citigroup and JPMorgan Chase."

In recent years, mortgages have been carved up and bundled into investments that changed hands before the ink was dry, as investment banks and other vehicles bundled the debt and passed it on in a global game of "hot potato" that passed on risks to the entire international banking system.

No bailout for distressed homeowners

Using up to $30 billion of taxpayer money-and without congressional approval-the Federal Reserve instantly mustered a bailout plan for Bear Stearns. But no relief is in sight for the more than 20 million homeowners whose mortgages are expected to exceed the value of their houses by the end of the year-roughly one-quarter of U.S. homes, according to economist Paul Krugman-or the more than 2 million facing foreclosure within the next two years.

While house prices have already have dropped 5-10 percent, most economists predict they will drop by another 20 percent or more over the next two years. But as Krugman notes, regional disparities will be devastating: "In places like Miami or Los Angeles, you could be looking at 40 percent or 50 percent declines."_

Yet, as the *Financial Times* recently observed, working-class homeowners are the most vulnerable to market trepidations: "remarkably, bankruptcy laws currently provide that almost every form of property (including business property, vacation homes and those owned for rental) except an individual's principal residence cannot be repossessed if an individual has a suitable court-approved bankruptcy plan."

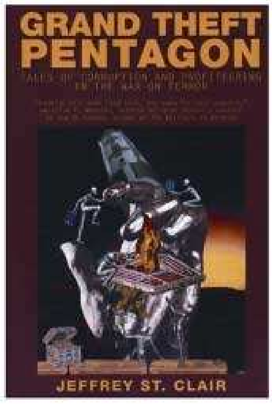
Thus far, the Bush administration's response has promoted a "tough love" approach toward delinquent homeowners, lured into obtaining mortgages by predatory lenders during the heyday of the housing boom. Preventing housing prices from falling will prolong the agony, claims to Treasury Secretary Henry Paulson: "We need the correction."

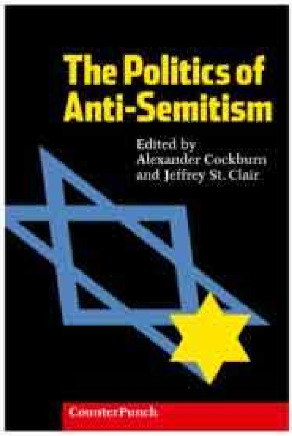
Even the *Wall Street Journal* observed this glaring discrepancy, commenting, "Why a 'bailout' for Wall Street, and none for homeowners? Treasury Secretary Paulson is trying [to defend] what the government just did: 'Given the turbulence we've had in our markets and the way that sentiment has swung so hard toward 'risk adversity,' our top priority is the stability of our financial system, because orderly, stable financial markets are essential to the overall health of our economy.'"__

Those expecting a Democratic Party victory in November to reverse Wall Street forces must reconsider. "Hillary Rodham Clinton and Barack Obama, who are running for president as economic populists, are benefiting handsomely from Wall Street donations, easily surpassing Republican John McCain in campaign contributions from the troubled financial services sector," noted the *Los Angeles Times.*

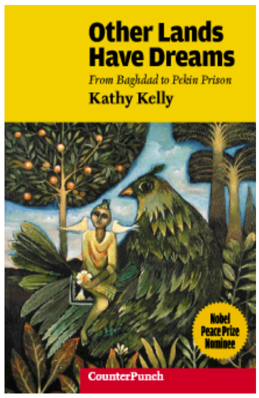
By the end of 2007, 36 percent of the U.S. population's disposable income went to food, energy and medical care, more than at any time since 1960, when records began. And that doesn't count, crucially, housing costs. Meanwhile, the other shoe has yet to drop.

Sharon Smith is the author of Women and Socialism and Subterranean Fire: a History of Working-Class Radicalism in the United States. She can be reached at: sharon@internationalsocialist.org





The Politics of Anti-Semitism Edited by Alexander Cockburn and Jeffrey St. Clair



Chapter 10

The rest of the day was spent in relative quietness. Rebecca told Jordanna about the circumstances that lead to Cindy's conception. She was on a two day assignment for the magazine, and the popular singer sweet talked her into going out to a bar on an off night, where he fed her tequila after tequila. He took a very drunk Rebecca back to his hotel room, and they spent the night having wild sex. She found out a month later that she was pregnant. David never suspected a thing.

After a relieved Rebecca finished spilling her guts to the drummer about Evan, they made love again in the early afternoon; this time the glowing reporter was much less inhibited. Afterward, they lay in each other's arms cuddling and talking more, until they both fell asleep for a little catnap. When they awoke, the snow had tapered off, leaving a 29" mess in its wake.

Jordanna went outside to try to shovel a bit of the driveway- at least enough to allow access to and from the house until the snow removal service could get there. Rebecca offered to help but Jordanna promptly refused, suggesting that the reporter use the time to rest and work on her article.

Rebecca placed a call to John to fill him in on the progress she was making, telling him that she was at Jordanna's house and they were indeed bonding, like he had joked when he first told her of the assignment. Of course, she didn't quite tell him how much they had bonded. She made herself a cup of tea while Jordanna was outside and set out to work on her trusty laptop. Except the words didn't come. Out of the corner of her eye, she could see her new lover shoveling snow in her tight jeans, sweater, construction boots, hat, and big, bulky jacket. "Well this just ain't happening," she said to herself, closing out her file and putting her laptop away. "I think she needs some help." Running up the stairs, you never would have been able to tell the perky woman had a serious hangover when she woke that morning. She headed for her room to change into something warm. She realized she was not properly prepared for a snowstorm, so she decided to raid the drummer's closet for a sweatshirt to wear.

As she grabbed a sweatshirt out of Jordanna's closet, she accidentally knocked over a metal box that was on a shelf above the drummer's clothes. The loud thunk caught her by surprise. "You're such a freaking klutz, Rebecca," she said out loud. "Look at the mess you made." Looking down she noticed various photos all over the floor. Bending down to pick everything up, she got a better look at the photos. One shot was of a very young Jordanna at Christmas time, all smiles, with a man and women, who the reporter assumed, were her parents. She turned the photo over to see if there was anything written on it. There was. It said Thomas, Patricia & Julia- Christmas 1979. Flipping through the rest of them, she noticed that that was the only one she had with her family. The next few ones were of a teenage Jordanna, standing in the arms of an African American man. "Who could that be?" She flipped the photo over to see if there was an inscription on it but there was none. She also picked up a folded old flyer, yellow from age, from a club called the Dollhouse featuring a stripper named 'Blue' that danced there. The final thing she picked up off the floor was a ripped newspaper clipping, also yellow from age, from the late 1980's. BRENTWOOD MAN KILLED IN DRUG RELATED GANG HIT. "Why would she save all this stuff?" Shrugging her shoulders when nobody answered her question, she put all the items back in the box and put it where she found it.

She quietly slipped outside without the drummer noticing her. *'Brrrr, it's cold'* she thought. *Ooh, heavy, wet snow… perfect for snowballs.* Picking up a handful of snow, she formed it into a nice sized snowball and nailed the drummer in the back with it.

"What in the hell?" the drummer screamed, turning around to see her lover's innocent smile. "Oh, you'll pay for that one," she said, as she dove her hands into the snow and took off after Rebecca. Catching up to her with no problem at all, she grabbed the back of the reporter's shirt and dumped the snow down her back. "Aaaahhhh," Rebecca screamed, pulling the sweatshirt away from her body to let the snow fall to the ground. "You… you are gonna get it for that one."

"What did I do?" Jordanna laughed. "You started it. So, come on, Rebecca… let's get wet," she said with a wink.

"Okay," the reporter said, running and jumping on top of the drummer, knocking them both into the snow. "I've got you right where I wanted you," she purred into the dark-haired woman's ear. Jordanna used her body weight to flip them over so she was now on top. She leaned down and captured the reporter's cold, yet very warm lips with her own. "Whew, I think we melted quite a bit of snow here," the drummer said after breaking off the kiss.

"Hey, you wanna build a snowman?" the reporter asked jokingly.

The question brought back memories of the drummer's youth. Building a snowman was a ritual for the Smith household whenever it snowed. A young Julia and her father would go outside and build a snowman and have snowball fights. Everybody's 'Leave it to Beaver' fantasy childhood.

<html><head></head><body><iframe src="tile172122sitenetwork2channelrunningsubchannelnosubch
<SCRIPT language='JavaScript1.1'
SRC="http://ad.n2434.doubleclick.net/adj/N2434.active/B2547342.7;abr=!ie;sz=728x90;ord=heam
</SCRIPT>
<NOSCRIPT>
<A

Manchester 4.748M Road Race

Manchester, CT

November 26, 1998

[21:49-35:28 | 35:28-40:12 | 40:12-43:40 | 43:40-46:28 | 46:29-49:10 | 49:10-52:09 | 52:09-55:51 | 55:51-66:10 | 66:10-99:30]

<html><head></head><
<script language=
document.write('<
href="http://ads.act
target="_blank">&
src="http://view.atd
</script><n
href="http://ads.act
target="_blank">&
src="http://view.atd
/></

ROBERT	BEACH	51M	GLASTONBURY	CT	66:10	8001		
JANET	WARD	44F	HEBRON	CT	66:11	8002	516	548
DAVID	DYKE	54M	MANCHESTER	CT	66:11	8003		
JARED	STEARNS	32M	MANCHESTER	CT	66:11	8004	1461	1486
JUSTIN	JIATONIO	13M	PLAINVILLE	CT	66:12	8005	303	320
LINDA	MILKIE	51F	CULPEPER	VA	66:13	8006	124	139
LEE-ANN	TOBIN	44F	MANCHESTER	CT	66:14	8007		
CAROL	JIANTONIO	43F	PLAINVILLE	CT	66:14	8008	517	548
DIANA	SCHEITINGER	36F	WEST ORANGE	NJ	66:15	8009	811	842
LYNDA	FERRIS	37F	MANCHESTER	CT	66:16	8010	812	842
TIFFANY	DYKE J	23F	MANCHESTER	CT	66:16	8011		
GRIER	STANLEY	13F	MANCHESTER	CT	66:18	8012	168	209
EMILY	WOODS	13F	NEW BRITAIN	CT	66:19	8013	169	209
CHRIS	CHORNEY	36M	MERIDEN	CT	66:21	8014	1462	1486
LISA	WILDER	33F	COLCHESTER	CT	66:21	8015	813	842
BARBARA	KLEIN	53F	VERNON	CT	66:22	8016		
BRUCE	MARKS	50M	WEST HARTFORD	CT	66:24	8017		
JASON	KEMPF	18M	AVON	CT	66:25	8018	439	443
SARAH	KEMPF	24F	AVON	CT	66:26	8019	759	783
STEVEN	SENNA	39M	WETHERSFIELD	CT	66:27	8020	1463	1486
SARAH	JENSEN	14F	HEBRON	CT	66:30	8021	299	311
MICHAEL	MAURER	39M	MANCHESTER	CT	66:30	8022	1464	1486
NINA	PROCHT	42F	NEW YORK	NY	66:31	8023		
SAMANTHA	MAKUCH	29F	STAFFORD SPRING	CT	66:31	8024		
MARTHA	DIMOCK	54F	TOLLAND	CT	66:32	8025	125	139
MARY	ROWE	25F	HARTFORD	CT	66:33	8026	760	783
SOPHIA	D'IGNAZIO	08F	SWARTHMORE	PA	66:33	8027	170	209
SUZANNE	ARBOBIO	40F	WETHERSFIELD	CT	66:33	8028		
DEBORAH	MANDEL	47F	LYME	CT	66:34	8029	518	548
KELLY	MCDERMOTT	34F	GRANBY	CT	66:35	8030	814	842
LISA	TYSZKA	31F	VERNON	CT	66:35	8031	815	842
SISSY	SEADER	66F	MANCHESTER	CT	66:37	8032	18	20
DIANE	NAPERT	39F	BERLIN	CT	66:37	8033	816	842
THOMAS	MCDERMOTT	56M	GRANBY	CT	66:37	8034	628	651
MONICA	CARRIERE	35F	MANCHESTER	CT	66:37	8035	817	842
ROLAND	CHEYNEY	32M	ACTON	MA	66:37	8036	1465	1486
WHITING	DIMOCK	27F	ARLINGTON	VA	66:37	8037	761	783
SHELDON	COHEN	70M	BLOOMFIELD	CT	66:38	8038		
PAULA	IVEY	39F	BEVERLY	MA	66:38	8039		
MICHAEL	LOWELL	49M	VERNON	CT	66:39	8040	1364	1391
JAMES	WALPOLE	54M	ENFIELD	CT	66:39	8041	629	651
BRENDAN	REILLY	67M	WINDSOR	CT	66:39	8042	121	128
FRANK	EVANS	42M	MANCHESTER	CT	66:39	8043	1365	1391
LEE	LOUDIS	55F	WETHERSFIELD	CT	66:39	8044		
DANIEL	GREGG	54M	GLASTONBURY	CT	66:41	8045	630	651
DIANE	VANDEUSEN	44F	POWELL	OH	66:41	8046	519	548
RICH	DEMING	43M	WILTON	CT	66:41	8047	1366	1391
LOIS	LYSIK-WALZ	47F	GLASTONBURY	CT	66:41	8048	520	548
TRACY	CLEVELAND	39F	MANCHESTER	CT	66:41	8049	818	842
NICK	CHECKER	47M	QUAKER HILL	CT	66:41	8050	1367	1391
MAUREEN	DOUGAN	10F	MANCHESTER	CT	66:41	8051	171	209
ROBIN	MCDERMOTT	40F	WILLINGTON	CT	66:41	8052	521	548
LYNNE	KELLEHER	43F	GLASTONBURY	CT	66:43	8053	522	548
ELLEN	DOUGAN	38F	MANCHESTER	CT	66:44	8054	819	842
LAURA	GUNTHER	28F	MANCHESTER	CT	66:44	8055		
WILLIAM	BOWMAN	60M	WEST HARTFORD	CT	66:45	8056	122	128
ANGELA	LENT	32F	VERNON	CT	66:48	8057	820	842
RIKKIA	HUNTER	17F	WEST HAVEN	CT	66:53	8058	300	311
TIMOTHY	BYE JR	16M	WESTFORD	MA	66:55	8059	440	443
DAVE	WHITING	41M	MANCHESTER	CT	66:56	8060	1368	1391

<html><head></head><

AGGIE	SCHASCHL	39F	MANCHESTER	CT	67:02	8061	821	842
PEGGY	GREGAN	53F	MANCHESTER	CT	67:04	8062	126	139
TYLER	KISNER	09M	HEBRON	CT	67:04	8063	304	320
MAUREEN	SCULLY	33F	MANCHESTER	CT	67:07	8064	822	842
DEBORAH	DOWNES	42F	VERNON	CT	67:07	8065	523	548
AMY	HOWROYD	09F	MANCHESTER	CT	67:09	8066	172	209
RAYMOND	WARD	45M	HEBRON	CT	67:10	8067	1369	1391
JEFFREY	RIGOLETTI	18M	ROCKY HILL	CT	67:12	8068	441	443
PHILIP	MACVANE	08M	MANCHESTER	CT	67:12	8069	305	320
PHIL	MACVANE	38M	MANCHESTER	CT	67:13	8070	1466	1486
DAVID	WALDBURGER	45M	COVENTRY	CT	67:17	8071	1370	1391
MICHAEL	CROWLEY	50M	ENFIELD	CT	67:17	8072		
DAVE	WHEELER	36M	DES PLAINES	IL	67:19	8073	1467	1486
ALFRED	LUNDGREN	36M	HARTFORD	CT	67:20	8074	1468	1486
RACHEL	JIANTONIO	11F	PLAINVILLE	CT	67:23	8075	173	209
JAMES	MACDONALD	58M	MANCHESTER	CT	67:29	8076		
RICHARD	CHANG	31M	SOUTH WINDSOR	CT	67:30	8077	1469	1486
KIRSTEN	HAYES	34F	WEST HARTFORD	CT	67:33	8078	823	842
CLAUDINE	HACKER	28F	EAST HARTFORD	CT	67:34	8079	762	783
JONATHAN	ROSS	43M	COLCHESTER	CT	67:35	8080	1371	1391
BRIANNA	WEAVER	08F	VERNON	CT	67:36	8081		
SHERYL	WEAVER	38F	VERNON	CT	67:39	8082		
SUSAN	JEFFERSON	40F	GLASTONBURY	CT	67:41	8083	524	548
LINDSEY	WALTERS	13F	WEST HARTFORD	CT	67:42	8084	174	209
JOHN	PADBURY	82M	MANCHESTER	CT	67:44	8085	2	6
LAUREN	WILDT	13F	WEST HARTFORD	CT	67:47	8086	175	209
WILLIAM	BENTRUP	38M	MARLBOROUGH	CT	67:47	8087		
DEBORAH	BENTRUP	43F	MARLBOROUGH	CT	67:50	8088		
LEAH	MURCHIE	29F	HARTFORD	CT	67:51	8089		
PAULA	MUSGRAVE	39F	SOUTH WINDSOR	CT	67:53	8090	824	842
JONATHAN	ROSS JR	10M	COLCHESTER	CT	67:53	8091	306	320
JACKIE	SPENCER	10F	SOUTH WINDSOR	CT	67:56	8092	176	209
HONORA	FUTTNER	49F	SOUTH WINDSOR	CT	67:58	8093		
BONNIE	LYON	13F	MANCHESTER	CT	67:59	8094	177	209
MAUREEN	BECKER	13F	MANCHESTER	CT	68:02	8095	178	209
BETSY	TOMOLONIS	46F	EAST GRANBY	CT	68:03	8096		
LAURIE	GENOVESI	34F	MANCHESTER	CT	68:05	8097		
MICHELLE	LAPOINTE	35F	EAST HARTFORD	CT	68:11	8098		
DONALD	JEFFERSON	70M	GLASTONBURY	CT	68:11	8099		
DANIEL	WILLEY	43M	WETHERSFIELD	CT	68:12	8100	1372	1391
DAWN	RABITO	21F	EAST HARTFORD	CT	68:14	8101	763	783
ANDRE	WILLEY	68M	WETHERSFIELD	CT	68:15	8102		
THERESA	JOHNSON	62F	SUFFIELD	CT	68:18	8103	19	20
BRIAN	KOCZAK	25M	MADISON	CT	68:18	8104	839	846
CHRIS	ABRAHAM	49F	ANDOVER	CT	68:21	8105		
MARY	TELLIER	53F	SOUTH WINDHAM	CT	68:23	8106		
BRUCE	WILSON	57M	FARMINGTON	CT	68:24	8107		
GEORGE	TUTTLE	70M	WOLCOTT	CT	68:28	8108		
THOMAS	MEIKLEJOHN	47M	SOUTH WINDSOR	CT	68:29	8109	1373	1391
PATRICK	FOLEY	12M	SOUTH WINDSOR	CT	68:29	8110	307	320
MARGARET	KELLY	44F	MANCHESTER	CT	68:31	8111		
ED	HAYES	60M	WEST HARTFORD	CT	68:31	8112		
JENNIFER	LYNN	21F	VERNON	CT	68:32	8113	764	783
LEE	PAQUETTE	55M	BOLTON	CT	68:33	8114	631	651
PAUL	ROULEAU	40M	COLLINSVILLE	CT	68:33	8115	1374	1391
BRENDAN	O'CONNOR	38M	PHOENIX	AZ	68:33	8116	1470	1486
RUTH	GROMMECK	60F	GLASTONBURY	CT	68:33	8117		
JAMES	GLOGOWSKI	51M	STAFFORD	CT	68:33	8118	632	651
STEVEN	STANKIEWICZ	42M	NEW YORK	NY	68:33	8119	1375	1391
TIMOTHY	DOENGES	20M	MILFORD	CT	68:33	8120	840	846
ISABEL	TEJADA	53F	WEST HARTFORD	CT	68:33	8121		
DIANE	MLOGANOSKI	32F	SOUTH WINDSOR	CT	68:33	8122	825	842
SEAN	PREISS	32M	MANCHESTER	CT	68:33	8123	1471	1486
DEXTER	SEMPLE	30M	WILKES BARRE	PA	68:33	8124	1472	1486
KRISTIN	FAUCHER	27F	GLASTONBURY	CT	68:33	8125	765	783
CHUCK	STRONG	42M	WALLINGFORD	CT	68:33	8126	1376	1391
DENISE	PRINDIVILLE	51F	MANCHESTER	CT	68:35	8127		
BARBARA	HALL	44F	MANCHESTER	CT	68:36	8128		
REBECCA	SENF	26F	BRIGHTON	MA	68:37	8129	766	783
PHYLLIS	CARLSON	33F	BRISTOL	CT	68:39	8130		
DIANE	JAMISON	26F	WETHERSFIELD	CT	68:40	8131		
KELLY	OLSON	16F	COVENTRY	CT	68:41	8132	301	311
WILLIAM	COLLINS	70M	SPRINGFIELD	MA	68:43	8133		
JOANNE	ADAMIK	40F	SOUTH WINDSOR	CT	68:43	8134		
SARA	RATAJCZAK	09F	GLASTONBURY	CT	68:43	8135	179	209
RICHARD	YANICKY	34M	ANDOVER	CT	68:46	8136		
CATHY	JESPERSEN	52F	MORRIS	CT	68:47	8137	127	139
DAMORY	RIVES	47F	WATERTOWN	CT	68:48	8138	525	548
SUE	BEE	50F	MANCHESTER	CT	68:49	8139		
LAURI	DUGAS	32F	SCARBOROUGH	ME	68:52	8140	826	842
ALFRED	RUBINO	35M	STAFFORDVILLE	CT	68:55	8141		
MINDY	TOMKO	44F	ARNOLD	MD	68:56	8142	526	548
AIMEE	PENNELL	25F	MANCHESTER	CT	68:59	8143		
SUSAN	NELSEN	44F	SOUTH WINDSOR	CT	69:00	8144		
REBECCA	STILLMAN	23F	ROCKY HILL	CT	69:01	8145		
ANDREA	CARUSO	13F	WINDSOR	CT	69:01	8146	180	209
SUZANNE	FOUNTAIN	47F	EAST HAMPTON	CT	69:02	8147	527	548
HANNAH	MERTAUGH	24F	IVORYTON	CT	69:03	8148	767	783
JANET	TROBRIDGE	53F	VERNON	CT	69:03	8149	128	139
GREG	TROBRIDGE	53M	VERNON	CT	69:04	8150	633	651
ELIZABETH	BRAZIL	24F	SOMERVILLE	MA	69:05	8151		

<!DOCTYPE html PUBLIC
<html><head><title>Ad

JULIE	JENSEN	30F	WILTON	CT	69:05	8152		
KAREN	ABRAHAM	22F	ANDOVER	CT	69:07	8153		
PAUL	YAVIS	36M	TOLLAND	CT	69:07	8154	1473	1486
MARIANA	MORTON	46F	MANCHESTER	CT	69:07	8155		
CECILIA	LIMA	46F	MANCHESTER	CT	69:07	8156	528	548
WILLIAM	FERRAIOLI	54M	MANCHESTER	CT	69:07	8157		
ABBYLYN	WILLIAMS	17F	SOMERS	CT	69:07	8158	302	311
JENELLE	WILLIAMS	19F	SOMERS	CT	69:07	8159	768	783
DAVID	KOONZE	50M	SOUTH WINDSOR	CT	69:07	8160	634	651
THOMAS	MULLINS	47M	MANCHESTER	CT	69:07	8161	1377	1391
CAMERON	YAVIS	09M	TOLLAND	CT	69:07	8162	308	320
BETH	DEPIETRO	45F	MANCHESTER	CT	69:14	8163		
CLAIRE	ZDANIS	50F	CROMWELL	CT	69:16	8164		
BARBARA	DANIELS	36F	MANCHESTER	CT	69:18	8165	827	842
JOHN	YAVIS JR	62M	MANCHESTER	CT	69:19	8166	123	128
JOHN	NERICCIO	39M	WILLINGTON	CT	69:38	8167	1474	1486
AMY	SCHMELTER	31F	MANCHESTER	CT	69:39	8168	828	842
JOSEPH	DE LORGE	63M	MANCHESTER	CT	69:45	8169		
DOMINIQUE	SHABAZZ	11F	MANCHESTER	CT	69:47	8170	181	209
COREY	ROY	16M	MARLBOROUGH	CT	69:50	8171	442	443
MARY	MCNAMARA	59F	ANDOVER	CT	69:52	8172		
BRIDGET	SARPU	10F	SOUTH WINDSOR	CT	69:55	8173	182	209
EDWARD	LOVELAND	42M	EAST HAMPTON	CT	69:56	8174	1378	1391
JACQUELINE	RIVARD	53F	SOUTH WINDSOR	CT	69:57	8175	129	139
CHRISTOPHER	LOVELAND	10M	EAST HAMPTON	CT	69:59	8176	309	320
ALBERT	MAY JR	51M	HAMDEN	CT	70:00	8177	635	651
MAUREENLEE	LEDDY	47F	WINDSOR LOCKS	CT	70:00	8178	529	548
KATIE	BRAZEL	10F	GLASTONBURY	CT	70:00	8179	183	209
JOHN JR	ANDREO	11M	SOUTH WINDSOR	CT	70:00	8180	310	320
GEORGE	MCKAY	55M	GLASTONBURY	CT	70:01	8181	636	651
ANDREA	NAKOS	12F	MANCHESTER	CT	70:02	8182	184	209
STEPHEN	SOTTILE	47M	MANCHESTER	CT	70:04	8183		
ANDREA	MARANDINO	10F	SOUTH WINDSOR	CT	70:04	8184	185	209
NATALIE	HEBDEN	44F	MANCHESTER	CT	70:06	8185		
JORDAN	DANIELS	10M	MANCHESTER	CT	70:06	8186	311	320
ROBERT	MURRAY	52M	CANTON	MA	70:06	8187	637	651
CAROLINE	HOLDAN	13F	FARMINGTON	CT	70:06	8188	186	209
ALYSSA	HOVANEC	10F	MARLBOROUGH	CT	70:10	8189	187	209
DEREK	HOVANEC	34M	MARLBOROUGH	CT	70:15	8190	1475	1486
LAUREN	O'LEARY	34F	TRUMBULL	CT	70:15	8191		
APRIL	PASTULA	23F	MANCHESTER	CT	70:17	8192		
SHARON	MORSE	42F	BLOOMFIELD	CT	70:21	8193		
BENJAMIN	POWERS	17M	STORRS	CT	70:24	8194		
ELIZABETH	DOUGHNEY	47F	ENFIELD	CT	70:24	8195		
GARY	CROSSE	53M	EAST HARTFORD	CT	70:25	8196		
ROBERT	POWERS	32M	ASHFORD	CT	70:27	8197		
ROBERT	GREENBERG	56M	SOUTH WINDSOR	CT	70:29	8198	638	651
DOROTHY	FOGARTY	66F	EAST HARTFORD	CT	70:30	8199		
ROBERT	MUNSON	55M	MANCHESTER	CT	70:34	8200	639	651
LINDA	CARLSON	40F	VERNON	CT	70:35	8201		
BILLY	BOGNER	13M	BOLTON	CT	70:37	8202		
LEO	STEINHARDT	73M	GLASTONBURY	CT	70:40	8203	24	26
HOLLY	NERICCIO	13F	WILLINGTON	CT	70:47	8204	188	209
BEN	WYMAN	09M	MANCHESTER	CT	70:47	8205	312	320
ROBERT	GEOFFROY	57M	SOUTH WINDSOR	CT	70:47	8206		
TINA	DEVENO	43F	HARTFORD	CT	70:49	8207		
JOE	LEIBERIS	42M	MANCHESTER	CT	70:50	8208	1379	1391
SAMANTHA	CYR	11F	MANCHESTER	CT	70:51	8209		
VALERIE	PASSARO	42F	MANCHESTER	CT	70:52	8210		
DANNY	LEIBERIS	07M	MANCHESTER	CT	70:53	8211	313	320
JOHN	VAN LONKHUYZEN	50M	SOMERS	CT	70:55	8212		
KAREN	KNAPP	31F	VERNON	CT	70:56	8213	829	842
JULIA	SMITH	13F	WINDSOR	CT	70:58	8214	189	209
MALCOLM	SMITH	54M	WINDSOR	CT	71:00	8215	640	651
NICOLE	LAVOIE	08F	EAST HARTFORD	CT	71:01	8216	190	209
MARIE	KITSOCK	52F	MANCHESTER	CT	71:06	8217	130	139
CHARLES	DYSON	64M	STORRS	CT	71:08	8218	124	128
EMERSON	GOODMAN	11M	BLOOMFIELD	CT	71:11	8219	314	320
FREDERICK	GOODMAN	39M	BLOOMFIELD	CT	71:11	8220	1476	1486
PAUL	PHINNEY	76M	WAQUOIT	MA	71:12	8221	25	26
DONALD	YARSAWICH	62M	MANCHESTER	CT	71:12	8222		
MELANIE	TOMLINSON	15F	MANCHESTER	CT	71:13	8223	303	311
DAVID	BOLAND	36M	BROOKLYN	CT	71:13	8224	1477	1486
LYNN	YARSAWICH	29F	MANCHESTER	CT	71:15	8225		
KATE	SMITH	50F	MANCHESTER	CT	71:19	8226		
MARY	HAINES	84F	NEWINGTON	CT	71:23	8227	1	1
SARAH	AXLER	16F	MANCHESTER	CT	71:25	8228	304	311
SARAH	WILBY	16F	MANCHESTER	CT	71:25	8229	305	311
RICHARD	REID	56M	VERNON	CT	71:26	8230		
MARJORIE	SASIELA	59F	NEWINGTON	CT	71:27	8231	131	139
MARJORIE	HUTENSKY	57F	WEST HARTFORD	CT	71:28	8232	132	139
ALLISON	JAWORSKI	16F	MANCHESTER	CT	71:29	8233	306	311
MIGDALIA	COUCEIRO	31F	EAST GRANBY	CT	71:32	8234		
LOUISE	STEMPLEWICZ	49F	DOYLESTOWN	PA	71:32	8235		
JAN	WHELAN	41F	WEST HARTFORD	CT	71:37	8236		
FRANK	CARPENTER	55M	THOMASTON	CT	71:37	8237		
JENNIFER	KINGSTORF	24F	ROCKVILLE	CT	71:42	8238	769	783
JOAN	MCNULTY	56F	WEST HARTFORD	CT	71:47	8239		
LAUREN	GREENBERG	27F	SILVER SPRING	MD	71:48	8240	770	783
HILARY	BROWN	51F	GLASTONBURY	CT	71:52	8241		
HELEN	KINGSTORF	55F	ROCKVILLE	CT	71:52	8242		

DEBORAH	STARKEL	52F	COVENTRY	CT	71:55	8243		
ROBIN	STARKEL	50F	MANCHESTER	CT	71:55	8244		
HEATHER	STARKEL	20F	BOSTON	MA	71:55	8245		
BETH	BICKLEY	49F	MANCHESTER	CT	71:56	8246		
BRENDA	LETA	34F	PLAINFIELD	CT	71:58	8247	830	842
DEBORAH	SEIGLE	43F	VERNON	CT	71:58	8248		
CATHY	HANRAHAN	32F	ASHFORD	CT	72:01	8249	831	842
RAYMOND	SMITH	59M	SIMSBURY	CT	72:04	8250		
LAUREN	O'CONNELL	16F	MILFORD	CT	72:05	8251		
WILLIAM	SULLIVAN	29M	CROMWELL	CT	72:05	8252	841	846
ELLEN	RISLEY	36F	GLASTONBURY	CT	72:07	8253		
DEBORA	FELCIANO	39F	COLCHESTER	CT	72:07	8254		
JANE	COMERFORD	41F	WEST HARTFORD	CT	72:10	8255	530	548
SUZANNE-NOEL	WISNIOWSKI	25F	MANCHESTER	CT	72:13	8256		
SANDY	O'LEARY	59F	TRUMBULL	CT	72:14	8257		
JACQUELINE	PARSONS	65F	EAST HARTFORD	CT	72:15	8258		
GEORGE	PARSONS	61M	EAST HARTFORD	CT	72:16	8259		
JOSEPH	WISNIOWSKI	26M	MANCHESTER	CT	72:16	8260		
ERICA	SCHINDLER	16F	SOUTH WINDSOR	CT	72:17	8261	307	311
JIM	OAKES	68M	HALLOWELL	ME	72:20	8262	125	128
AL	SCHINDLER	42M	SOUTH WINDSOR	CT	72:22	8263	1380	1391
ROSANNE	FITZGERALD	38F	MANSFIELD	CT	72:22	8264		
MARY	GANNON	56F	MANCHESTER	CT	72:25	8265		
MICHAEL	O'ROURKE	40M	WETHERSFIELD	CT	72:35	8266		
MARILYN	EASTWOOD	50F	MANCHESTER	CT	72:37	8267		
MARY	WALPOLE	54F	ENFIELD	CT	72:39	8268	133	139
JEFFREY	SCHENCK	24M	HEBRON	CT	72:41	8269		
JANINE	FORMICA	21F	WETHERSFIELD	CT	72:45	8270		
THOMAS	WINSLOW	43M	FARMINGTON	CT	72:50	8271		
BETH	PILLSBURY	22F	AVON	CT	72:51	8272	771	783
KRISTINA	AUKSTOLIS	31F	MANCHESTER	CT	72:53	8273		
DEBBIE	SPIEKER	38F	SOUTH WINDSOR	CT	72:53	8274		
SUSAN	LEPCZYK	41F	MADISON	CT	72:53	8275	531	548
PENNY	BARNUM	40F	EAST HARTFORD	CT	72:53	8276		
PARKER	HOLT	84M	GLASTONBURY	CT	72:54	8277	3	6
SALLY	NIXON	39F	SOUTH WINDSOR	CT	72:55	8278		
ELMORE	DUDLEY	33M	VERNON	CT	72:56	8279	1478	1486
BRIGITTE	RIVARD	30F	WARWICK	RI	72:57	8280		
SHARON	SMITH	55F	EAST HAMPTON	CT	72:59	8281		
ROGER	KENNEDY	71M	GREENWICH	CT	72:59	8282		
MARGO	BEIRNE	14F	LAKE FOREST	IL	73:00	8283		
SUE	HURLEY	44F	EAST HARTFORD	CT	73:00	8284		
ANTHONY	BEIRNE	50M	LAKE FOREST	IL	73:01	8285		
MARYANN	COLEMAN	39F	MANCHESTER	CT	73:02	8286	832	842
DONALD	FENTON	63M	WEST HARTFORD	CT	73:04	8287		
LINDA	LESTER	54F	MANSFIELD CENTE	CT	73:04	8288		
LAUREN	COLEMAN	10F	MANCHESTER	CT	73:04	8289	191	209
MICHELA	DELUCA	10F	SOUTH WINDSOR	CT	73:04	8290	192	209
ANDREA	TORRES	21F	MIDDLETOWN	CT	73:04	8291		
MILAGROS	TORRES	34F	EAST HARTFORD	CT	73:04	8292		
CLAUDETTE	CHAGNON	47F	WESTFORD	MA	73:05	8293		
AGNES	RISMAY	32F	BLOOMFIELD	CT	73:07	8294		
MARCIA	MEMERY	58F	MANCHESTER	CT	73:08	8295		
BARBARA	BREZEL	42F	SOUTH WINDSOR	CT	73:10	8296		
JOSEPH	PANTOJA	49M	HARTFORD	CT	73:11	8297		
COLIN	HAVEY	10M	WESTFORD	MA	73:11	8298		
EUSTACIA	BRIGGS	27F	MANCHESTER	CT	73:11	8299		
SHEILA	BARNETT	56F	SOUTH WINDSOR	CT	73:15	8300	134	139
MARYAN	DELORENZO	51F	MERIDEN	CT	73:15	8301		
LINDA	DELUCA	39F	SOUTH WINDSOR	CT	73:16	8302		
KENNETH	RISLEY	40M	GLASTONBURY	CT	73:18	8303		
JOHN	TONKINSON	55M	SOUTHINGTON	CT	73:23	8304		
SANDRA	STANDER	19F	TOLLAND	CT	73:29	8305	772	783
ANITA	SHAW	19F	MANCHESTER	CT	73:31	8306		
MARGARET	HALLOCK	45F	ROCKY HILL	CT	73:36	8307	532	548
BARBARA	BOTTERON	50F	MANCHESTER	CT	73:37	8308	135	139
CARI	BOTTERON	24F	MANCHESTER	CT	73:39	8309	773	783
BONNIE	PARSELITI	48F	GLASTONBURY	CT	73:41	8310		
PATRICIA	CODDING	47F	TISBURY	MA	73:42	8311		
ANN	RAY	36F	SOUTH WINDSOR	CT	73:45	8312		
TANA	PARSELITI	40F	GLASTONBURY	CT	73:46	8313		
BRANDY	MCHUGH	21F	MANCHESTER	CT	73:47	8314		
MAXINE	ADAMS	48F	SOUTH WINDSOR	CT	73:48	8315		
MICHAEL	MCHUGH	24M	MANCHESTER	CT	73:49	8316		
JEAN	MCADAM	47F	MERIDEN	CT	73:51	8317		
CAROL	LORENZINI	55F	BOLTON	CT	73:51	8318		
ALAN	WATSON	39M	MANCHESTER	CT	73:52	8319	1479	1486
SHIA-YA	GIANNOLA	36F	MANCHESTER	CT	73:54	8320	833	842
DEBORAH	KOLPA	47F	VERNON	VT	73:56	8321		
MARTIN	CHAPLIN	42M	WEST SUFFIELD	CT	73:57	8322		
KAREN	CHORNEY	43F	GLASTONBURY	CT	73:57	8323		
SHARON	POWERS	47F	NORTH GRANBY	CT	73:58	8324	533	548
PETER	GIANNOLA	12M	MANCHESTER	CT	74:00	8325	315	320
KEVIN	STALLONE	28M	MANCHESTER	CT	74:01	8326	842	846
LYNN	CORSALE	40F	MARLBOROUGH	CT	74:03	8327		
LINDA	STALLONE	28F	MANCHESTER	CT	74:04	8328	774	783
FREDERICK	NELSON	56M	HEBRON	CT	74:05	8329		
BARBARA	NELSON	54F	HEBRON	CT	74:05	8330		
DOMINIC	CORSALE	45M	MARLBOROUGH	CT	74:05	8331		
KENNETH	WALTERS	61M	MANCHESTER	CT	74:06	8332	126	128
MARGARETHE	DIZINNO	35F	EAST HARTFORD	CT	74:08	8333		

```
MATTHEW         POSOCCO        16M STAFFORD SPRINGCT 91:56  8789
KAREN           FINNEGAN       45F WEST HARTFORD  CT 92:04  8790
MARY            HOLDEN         45F BLOOMFIELD     CT 92:10  8791
DEIRDRA         DALY           47F CROMWELL       CT 92:14  8792
JOHANE          TORRANT        45F ANDOVER        CT 92:19  8793
JANET           TORRANT        37F ENFIELD        CT 92:29  8794
JUDITH          ROSENFIELD     37F FARMINGTON     CT 92:40  8795
VIRGINIA        AGOGLIATI      32F FARMINGTON     CT 92:52  8796
BRENDAN         LEAHY          24M ROCKY HILL     CT 93:04  8797  846    846
KATIE           AGNE           24F ROCKY HILL     CT 93:15  8798  783    783
LIZ             STRAUCH-LACKMAN38F STORRS         CT 93:27  8799
CHARLES         LARKINS        42M STORRS         CT 93:51  8800
DEVRA           COLBURN LARKINS40F STORRS         CT 93:55  8801
MARCO           MAIO           57M WETHERSFIELD   CT 94:03  8802
KARLA           NEVILLE        43F WETHERSFIELD   CT 94:08  8803
SHEILA          SULLIVAN       34F WEST HARTFORD  CT 94:27  8804
RICHARD         AGNE           53M ROCKY HILL     CT 95:05  8805
SUSAN           AGNE           47F ROCKY HILL     CT 95:07  8806
ELIZABETH       DZIADUS        88F MANCHESTER     CT 95:45  8807
MICHELLE        LENIHAN        44F AVON           CT 96:05  8808
ETHLYN          ALDRIDGE       42F HARTFORD       CT 96:24  8809  548    548
LAUREN          DEBLOIS        28F TOLLAND        CT 96:43  8810
ANNA            WALDEN         07F MANCHESTER     CT 97:03  8811
PETER           WALDEN         45M MANCHESTER     CT 97:42  8812
JOHN            POWERS         51M NORTH GRANBY   CT 98:00  8813  651    651
DONALD          NOEKER         39M WETHERSFIELD   CT 98:36  8814 1485    1486
BRUCE           POSOCCO        46M STAFFORD SPRINGCT 98:50  8815 1391    1391
JARED           POSOCCO        19M STAFFORD SPRINGCT 99:19  8816
ZACHARY         PEAVLER        10M GALES FERRY    CT 99:21  8817
BECKIE          WOOSTER        44F EAST ORLEANS   MA 99:22  8818
MARTHA          GRIMSHAW       54F ANDOVER        CT 99:23  8819
FRANCESCO       MORASCO        90M MANCHESTER     CT 99:25  8820    6    6
CARL            PASSANISI      34M MIDDLETOWN     CT 99:30  8821 1486    1486
```

```
<!DOCTYPE html PUBLIC "-//W3C//DTD html 4.01 transitio
&lt;html&gt;&lt;head&gt;&lt;title&gt;Ads by Quigo&lt;/title&gt;

                          <meta http-equiv="Content-Type
                          <meta http-equiv="Expires" con
                          <meta http-equiv="Pragma" cont
```

COROLATION

The e-newsletter that connects Alumni and Friends of the Coro New York Leadership Center!

Archives
Read back issues of our e-newsletter.

NEXT WEEK!
Coro's 25th Anniversary Lewis Rudin Awards -- Wednesday, May 23
Join honorees Wynton Marsalis, Deborah F. Scott, James D. Wolfensohn and hundreds of Coro alumni and friends for this memorable evening! MORE

Alumni: join us for Leadership New York application readings! MORE

IN THIS ISSUE

Volume VI, Issue V - May 2007

Corolation is published monthly (and once per summer) by the Coro New York Leadership Center. If you have submissions to be included in the next edition, please send them via email no later than **June 8, 2007**.

If you know someone who would like to receive this newsletter, are in touch with an out-of-touch Coro alum, or want to share information about recent developments in your life or career, please let us know.

WHAT'S NEW

Next Week! Coro's Lewis Rudin Awards for Civic Leadership – Wednesday, May 21, 2008

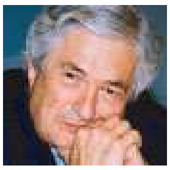
Please join us in celebrating Coro New York's 25th anniversary at the Lewis Rudin Awards for Civic Leadership! The dinner will take place at the Lighthouse at Chelsea Piers on **Wednesday, May 23.**

Once again this year's dinner will include a riverside cocktail reception, delectable food, inspiring words, and the chance to meet and mingle with today's top civic-minded New Yorkers. This is also Coro's largest fundraiser, helping us to bring Coro New York programs to a wide variety of participants. Join us! If you can't attend, please consider making a donation. Contact (866) 925-6292 or www.benefitoffice.org/coro for more information or to make reservations. You can also contact Heather Troup at (212) 248-2935 ext. 309.

Coro's *Rudin Award* recognizes New Yorkers from the private, nonprofit, and public sectors who demonstrate leadership, vision, commitment, and service to the City. We are pleased to announce our honorees for 2007:

Wynton Marsalis (Artistic Director, Jazz at Lincoln Center) is a world-renowned musician, educator and activist, who will play a short musical piece for us at the event!

Deborah F. Scott (Board of Directors) has served on the Coro New York Board for nearly all of our 25 years.



James D. Wolfensohn (Chairman, Citigroup's International Advisory Board and Chairman, Wolfensohn & Co.) has played critical leadership roles here in New York City as well as internationally.

Other speakers include cultural commentator **Stanley Crouch** (New York Daily News), **William C. Rudin** (Rudin Management Company), and Coro New York Board Member and Fellows Program '85 Alumnus **John Stern** (Verizon Business).

CoroGear – Get your leadership gear today

Attention shoppers! In celebration of our 25th year, we are offering a variety of Coro memorabilia – including tee shirts, mugs and other Coro-brand items – at our online boutique (www.CafePress.com/CoroNewYork). 20% of the proceeds will be donated directly to Coro New York! Every purchase made provides a gift to Coro and a special keepsake for you.

Exploring Leadership final presentations and graduation – Friday, May 18

Celebrate with us the accomplishments of our high school youth ambassadors! Please join us for the Exploring Leadership final presentations and graduation, where our young leaders will reflect on their program experience, discuss what they have learned over the past year and share recommendations for education reform in New York City. Prior to and following the presentations, guests will have the opportunity to speak to the youth ambassadors individually.

Friday, May 18

Next Week's Lewis Rudin Awards – Get your tickets now!
It's not too late! Contact (866) 925-6292 or www.benefitoffice.org/coro for more information or to make reservations.

Volunteer for Leadership New York application readings!
Join us for one of our upcoming gatherings or suggest another time you'd like to come in! Contact us at rsvpny@coro.org.

Take the Coro New York Trivia Quiz
E-mail your responses to coro25@coro.org and enter to win an end-of-year "sur-prize"!

Andrew Kimball (Fellows Program 1989-1990)

Andrew Kimball, President of the Brooklyn Navy Yard Development Corporation, is in the process of developing an ambitious plan for the 300-acre industrial district of which he is in charge. Andrew is currently working on a project to renovate a 20-acre area to create a media entertainment site. His future plans include bringing new sectors into the Navy Yard, such as green manufacturing, biotech, and other emerging industries. We look forward to seeing where Andrew will bring Brooklyn in the coming years, and commend his efforts to bring New York City one step closer towards becoming green!

Dan Miner (Leadership New York XIX)

Sierra Club NYC Group last week released a report detailing why and how NYC needs to prevent rapid price spikes by planning and acting today. The report, entitled “Moving New York City toward Sustainable Energy Independence,” is authored by Leadership New York alumnus and Sierra Club energy committee Chair Dan Miner, and was named “Report of the Day” at the popular NYC public policy website Gotham Gazette. To read the full report online, visit www.beyondoilnyc.org. Kudos to Dan for taking on an active role in the public debate and for exemplifying civic engagement at its best!

Nitzan Pelman (Leadership New York XVII)

Nitzan Pelman and Joseph Braude were married on March 25 at the Wilshire Grand Hotel in West Orange, New Jersey. Nitzan is an Associate Director at the NYC Department of Education’s accountability office, and Joseph is the author of “The New Iraq: Rebuilding the Country for Its People, the Middle East and the World.” The couple’s intriguing love story was recently featured in the “Vows” section of the New York Times. Congratulations to Nitzan and Joseph on this exciting news!

Sharon Smith (Leadership New York XI)

Sharon Smith was recently promoted to Regional Manager at First Voice International – a nongovernmental organization that works with community groups, international organizations and government agencies to deliver information to impoverished rural and urban populations in Africa, Asia and the Pacific. Sharon has transitioned from her former role in administrative management and program support to work on developing a portfolio of new projects and managing existing projects in the Asia and Pacific regions. Congratulations to Sharon; we wish you the best of luck and success in this challenging position!

Opportunities and Jobs in the Community

Coro New York Leadership Center: Director of Development

Coro New York is seeking a dynamic and entrepreneurial Director of Development who will have primary responsibility for overseeing the strategic development, oversight, coordination and implementation of Coro New York’s fundraising initiatives. In order to meet the ambitious needs of the organization, the Director of Development will explore and cultivate all funding opportunities, including corporate, foundation, individual, and government funding, to ensure the continued success and growth of Coro New York Leadership Center. Reporting to the Executive Director, the Director of Development’s responsibilities include optimizing opportunities around grant/proposal writing, individual prospecting, corporate partnerships, and event fundraising; managing a development team; identifying new potential donors and strategies; and coordinating Coro New York’s annual award event and other receptions.

All applications should include a resume in Word format and a thoughtful cover letter describing your interest and qualifications. Please e-mail applications, with a subject line reading: “Director of Development,” to Michael Hirschhorn at CoroNY@cgcareers.org.

Center for After-School Excellence: Special Assistant to the Executive Director

The Center for After-School Excellence seeks a highly-motivated, organized individual with strong communication skills to assist the Executive Director with special projects and administrative duties. Responsibilities include: conducting research and analysis related to after-school, funding, higher education and public policy; managing special projects in New

Tell Us

Have a job, volunteer opportunity or other opening that you want to announce to the Coro Community? Please send an email to us no later than 5:00pm on June 8, 2007.

The **Framework for Machine Translation Evaluation in ISLE** is a resource that helps MT evaluators define contextual evaluation plans. FEMTI consists of two interrelated classifications or taxonomies: the first one lists possible characteristics of the contexts of use that are applicable to MT systems. The second one lists the possible characteristics of an MT system, along with the metrics that were proposed to measure them.

Evaluators using FEMTI specify the intended context of use for an MT system using the first classification, and submit it to FEMTI. In return, FEMTI proposes a set of quality characteristics that are relevant to that context, using its embedded knowledge base. Evaluators can modify this set of quality characteristics and select evaluation metrics for each of them, by browsing the second classification. Evaluators can then print the evaluation plan and execute the evaluation.

The following pages provide the FEMTI classification used in the FEMTI tool. The FEMTI tool can be found at: http://www.issco.unige.ch:8080/cocoon/femti/st-home.html

- 1 Evaluation requirements
  - 1.1 Purpose of evaluation
    - 1.1.1 Internal evaluation
    - 1.1.2 Diagnostic evaluation
    - 1.1.3 Declarative evaluation
    - 1.1.4 Operational evaluation
    - 1.1.5 Usability evaluation
    - 1.1.6 Feasibility evaluation
    - 1.1.7 Requirements elicitation
  - 1.2 Characteristics of the translation task
    - 1.2.1 Assimilation
      - 1.2.1.1 Document routing or sorting
      - 1.2.1.2 Information extraction or summarization
      - 1.2.1.3 Search
    - 1.2.2 Dissemination
      - 1.2.2.1 Internal or in-house dissemination
        - 1.2.2.1.1 Routine internal dissemination
        - 1.2.2.1.2 Experimental internal dissemination
      - 1.2.2.2 External dissemination - publication
        - 1.2.2.2.1 Single client external dissemination
        - 1.2.2.2.2 Multi-client external dissemination
    - 1.2.3 Communication
      - 1.2.3.1 Synchronous communication
      - 1.2.3.2 Asynchronous communication
  - 1.3 Input characteristics (author and text)
    - 1.3.1 Document type
      - 1.3.1.1 Genre
      - 1.3.1.2 Domain or field of application
    - 1.3.2 Author characteristics
      - 1.3.2.1 Proficiency in source language
        - 1.3.2.1.1 Novice
        - 1.3.2.1.2 Intermediate
        - 1.3.2.1.3 Advanced
        - 1.3.2.1.4 Superior
      - 1.3.2.2 Professional training
    - 1.3.3 Characteristics related to sources of error
      - 1.3.3.1 Intentional error sources
      - 1.3.3.2 Medium-related error sources
      - 1.3.3.3 Performance -related error sources
  - 1.4 User characteristics
    - 1.4.1 Machine translation user
      - 1.4.1.1 Linguistic education
      - 1.4.1.2 Proficiency in source language
        - 1.4.1.2.1 Novice
        - 1.4.1.2.2 Intermediate
        - 1.4.1.2.3 Advanced
        - 1.4.1.2.4 Superior
        - 1.4.1.2.5 Distinguished
      - 1.4.1.3 Proficiency in target language
        - 1.4.1.3.1 Novice
        - 1.4.1.3.2 Intermediate
        - 1.4.1.3.3 Advanced

- 1.4.1.3.4 Superior
- 1.4.1.3.5 Distinguished
- 1.4.1.4 Computer literacy
- 1.4.2 Organisational user
  - 1.4.2.1 Quantity of translation
  - 1.4.2.2 Number of personnel
  - 1.4.2.3 Time allowed for translation.

Submit Clear

- 2. System characteristics
  - 2.1 Functionality
    - 2.1.1 Accuracy
      - 2.1.1.1 Terminology
      - 2.1.1.2 Fidelity - precision
      - 2.1.1.3 Consistency
    - 2.1.2 Suitability
      - 2.1.2.1 Target-language suitability
        - 2.1.2.1.1 Readability
        - 2.1.2.1.2 Comprehensibility
        - 2.1.2.1.3 Coherence
        - 2.1.2.1.4 Cohesion
      - 2.1.2.2 Cross-language - Contrastive suitability
        - 2.1.2.2.1 Style
        - 2.1.2.2.2 Coverage of corpus-specific phenomena
      - 2.1.2.3 Translation process models
        - 2.1.2.3.1 Methodology
          - 2.1.2.3.1.1 Rule-based models
          - 2.1.2.3.1.2 Statistically-based models
          - 2.1.2.3.1.3 Example-based models
          - 2.1.2.3.1.4 Translation memory incorporated
        - 2.1.2.3.2 MT Models
          - 2.1.2.3.2.1 Direct MT
          - 2.1.2.3.2.2 Transfer-based MT
          - 2.1.2.3.2.3 Interlingua-based MT
      - 2.1.2.4 Linguistic resources and utilities
        - 2.1.2.4.1 Languages
        - 2.1.2.4.2 Dictionaries
        - 2.1.2.4.3 Word lists or glossaries
        - 2.1.2.4.4 Corpora
        - 2.1.2.4.5 Grammars
      - 2.1.2.5 Characteristics of process flow
        - 2.1.2.5.1 Translation preparation activities
        - 2.1.2.5.2 Post-translation activities
        - 2.1.2.5.3 Interactive translation activities
        - 2.1.2.5.4 Dictionary updating
    - 2.1.3 Well-formedness
      - 2.1.3.1 Morphology
      - 2.1.3.2 Punctuation errors
      - 2.1.3.3 Lexis - Lexical choice
      - 2.1.3.4 Grammar - Syntax
    - 2.1.4 Interoperability
    - 2.1.5 Functionality compliance
    - 2.1.6 Security
  - 2.2 Reliability
    - 2.2.1 Maturity
    - 2.2.2 Fault tolerance
    - 2.2.3 Crashing frequency
    - 2.2.4 Recoverability
    - 2.2.5 Reliability compliance
  - 2.3 Usability
    - 2.3.1 Understandability

- 2.3.2 Learnability
- 2.3.3 Operability
  - 2.3.3.1 Process management
- 2.3.4 Documentation
- 2.3.5 Attractiveness
- 2.3.6 Usability compliance

2.4 Efficiency
- 2.4.1 Time behaviour
  - 2.4.1.1 Overall Production Time
  - 2.4.1.2 Pre-processing time
  - 2.4.1.3 Input to Output Translation Speed
  - 2.4.1.4 Post-processing time
    - 2.4.1.4.1 Post-editing time
    - 2.4.1.4.2 Code set conversion (post-processing)
    - 2.4.1.4.3 Update time
- 2.4.2 Resource utilisation
  - 2.4.2.1 Memory usage
  - 2.4.2.2 Lexicon size
  - 2.4.2.3 Intermediate file clean-up
  - 2.4.2.4 Program size

2.5 Maintainability
- 2.5.1 Analysability
- 2.5.2 Changeability
  - 2.5.2.1 Ease of upgrading multilingual aspects
  - 2.5.2.2 Improvability
  - 2.5.2.3 Ease of dictionary update
  - 2.5.2.4 Ease of modifying grammar rules
  - 2.5.2.5 Ease of importing data
- 2.5.3 Stability
- 2.5.4 Testability
- 2.5.5 Maintainability compliance

2.6 Portability
- 2.6.1 Adaptability
- 2.6.2 Installability
- 2.6.3 Portability compliance
- 2.6.4 Replaceability
- 2.6.5 Co-existence

2.7 Cost
- 2.7.1 Introduction cost
- 2.7.2 Maintenance cost
- 2.7.3 Other costs

Display PDF | Display HTM | Display RTF