Features for Web Person Disambiguation

Paul Kalmar
Kalmar Research
17 Lake Helix Drive
La Mesa, CA 91941
1-619-460-6093
paul@KalmarResearch.com

Dayne Freitag
SRI International
3661 Valley Centre Drive
San Diego, CA 92130
1-619-379-1393
freitag@ai.sri.com

ABSTRACT
Entity disambiguation resolves the many to many correspondence between mentions of entities in text and unique real-world entities. Our entity disambiguation uses language-independent entity context to agglomeratively resolve mentions with similar names to unique entities. This paper describes our automatic entity disambiguation capability and assesses its performance on the second Web People Search task. This paper also introduces new features such as named entity list context.

Categories and Subject Descriptors
I.5.3 [Clustering]

General Terms
Algorithms

Keywords
Entity disambiguation, list context, web search.

1. INTRODUCTION
We use the term entity to mean a specific person or object. A mention is a reference to an entity such as a word or phrase in a document. Taken together, all mentions that refer to the same real world object model that entity (Mitchell et al. 2004). Entity disambiguation inherently involves resolving many-to-many relationships. Multiple distinct strings may refer to the same entity just as multiple identical mentions may refer to distinct entities (Bagga and Baldwin, 1998). Our entity disambiguation software is based largely on language-independent algorithms that resolve mentions in the context of the entire corpus. The system utilizes multiple types of context as evidence for determining whether two mentions correspond to the same entity and it automatically learns the weight of evidence of each context item via corpus statistics.

The goal of the Web People Search tasks (Artiles et al. 2007) (Artiles et al. 2009) is to assign Web pages to groups, where each group contains all (and only those) pages that refer to one unique entity. A page is assigned to multiple groups if it mentions multiple entities, for example “John F. Kennedy” and the “John F. Kennedy Library”. The pages were selected via a set of keyword queries, and the disambiguation is evaluated only on those query entities. This differs from our system in a few key ways: our system deals with mentions rather than documents, our system does not require a filter on mentions, and our system is generally used for large collections of documents containing very many names rather than small sets of highly ambiguous documents dealing with one specific name. Nevertheless, it was possible to run our entity disambiguation system on the Web People Search task data with almost no modifications and achieve accurate results.

The remaining sections of this paper describe our automatic entity disambiguation methodology, look into the utility of various features for different disambiguation tasks, and report on the performance of the system on the WePS 2 data.

2. METHOD
Although our main algorithm did not significantly change from the first iteration of WePS (Kalmar & Blume, 2007), our system was reengineered to allow for simple expansion to new features and new entity types. Below we will discuss our algorithm.

2.1 Dealing with Raw Web Data
The first challenge in dealing with data from the Web is to decide which documents are useful and what text from those documents contains relevant information. We copied the title element and converted all text chunks to paragraphs, eliminating all other HTML and script.

2.2 NER and Within-Document Disambiguation
When dealing with unstructured text, a named entity recognition (NER) system provides the input to the entity disambiguation. Our system for NER is the same one used in Kalmar & Blume (2007). As described in Blume (2005), the system next carries out entity type-specific parsing in order to extract entity attributes such as titles, generate standardized names (e.g. p_abdul_khan_p for "Dr. Abdul Q. Khan"), and populate the data structures (token hashes) that are used to perform the within-document entity disambiguation. We err on the side of not merging entities rather than incorrectly merging entities. Looking at multiple documents provides additional statistics. Thus, the cross-document disambiguation process described in the next sections will still merge some entities even within individual documents.

2.3. Feature Co-occurrence and Weighting
Our cross-document entity disambiguation relies on one key insight: an entity can be distinguished by the company it keeps. If Abdul Khan 1 associates with different people and organizations at different locations than Abdul Khan 2, then he is probably a different person. Furthermore, if it is possible to compare two entities based on one type of context, it is possible to compare them based on every type of context. Within each domain, we require a finite set of context items. We define a context item as a feature of a given type which co-occurs with the entity. Co-occurrence in most cases is restricted to features contained in the same document. Co-occurring locations, organizations, and persons are the standardized names derived in the entity information extraction phase of within-document disambiguation. We use the logarithm of the inverse name frequency (the number of standard person names with which this context item appears), INF, as a weight indicating the salience of each context item. Co-occurrence with a common name provides less indication that two

mentions correspond to the same entity than co-occurrence with an uncommon name.

2.4. Entity Comparison

Once feature sets are determined for each entity, we can iterate through each feature type and create a composite score comparing two entities.

We define a separate distance measure per feature type. We are able to discount the co-occurrence with multiple items as well as quantify an unexpected lack of shared co-occurrence by engineering each distance measure for each specific domain. The score produced by each distance measure may be loosely interpreted as the log of the likelihood of two randomly generated contexts sharing the observed degree of similarity.

In addition to the context-based distance measures, we utilize a lexical (string) distance measure based on exactly the same transformations as used to compare strings for intra-document entity disambiguation plus the Soundex algorithm (Knuth 1998) to measure whether two name tokens sound the same. A large negative score indicates a great deal of similarity (log likelihood).

Since we are working in an unsupervised manner, it is necessary to heuristically determine a threshold for declaring a match. We use twice the information of the most informative name token multiplied by a given weight (in general one, but can be varied to produce solutions with greater precision or recall). The intuition behind this threshold is that a person with the most rare first and last name should match with no other context. More common person names should require a greater amount of context to be declared the same person.

2.5. Blocking

To avoid the unnecessary and costly task of comparing every pair of mentions, we utilize a blocking mechanism and only compare entities that are likely to match. Blocking is an iterative process, going from small very specific blocks to large more general blocks. This allows entities which are very similar to have a chance to match first before being compared against entities which are less similar. The sequence of blocking that we used is as follows: name used in the WePS query, long form of name, normalized name using Soundex, and family name.

Within a block, all entities are compared against each other, greedily agglomerating any two entities which match.

3. FEATURES

Because our new system allowed us to easily manipulate features, we did a large amount of research into what features are helpful for disambiguating person named entities as opposed to other types of entities. In Section 3.3 we ask what utility various contextual feature types have for various entity types, and attempt to provide an empirical answer to this question. In the absence of such empirical evidence, however, we can look to the characteristics of the various disambiguation problems for clues concerning the appropriate use of features. A characteristic of word sense disambiguation that makes it different from name disambiguation is sensitivity to syntax. Although syntactic roles and placement might assist in the recognition of named entities, we do not expect syntax to be useful for their disambiguation, since all names of the same type are grammatically interchangeable.

One of the major factors in disambiguating named entities is name comparison. Small name constituents, such as middle initials or suffixes, can make a critical difference to correct disambiguation when aligned properly, even though they have little impact on string similarity.

Named entities are very important features for disambiguating other named entities, as associations between them are high and their frequency is low, relative to other types of words.

The feature set used in comparing two person entities in this iteration of WePS consists of the feature set used in the first iteration of WePS (Kalmar & Blume, 2007) with a few new additional features.

3.1. Features used in WePS 1

The features used are either features directly derived from a person or those derived from contextual information.

3.1.1. Person features from NER

Features derived from persons directly are the same features used in the first iteration of WePS, extracted in an identical manner. These features are: StandardName, TitleGender, Longname, Gender, Initial, NamePostModifier, NickName, NameToken, SoundexToken, TitleToken.

3.1.2. Contextual Named Entities

As in WePS 1, we use all named entities in a document, locations, organizations, and persons, as contextual features (although unlike in WePS1, these were used without a windowing method). Because many pages consisted mainly of lists of entities, a quota was put on documents to reduce noise and increase speed. A maximum of 300 of each entity type was taken from a document for use as context or for disambiguation. Ideally, the new list feature described in Section 3.2.3 will take care of these seemingly extraneous contextual items.

3.2. New features since WePS 1

Although there were many new features that we would have liked to experiment with adding, due to time constraints we only added URL, page title words, and lists.

3.2.1. URL

We have noticed that it is often the case that websites have multiple subpages that discuss the same person. For this reason, we found it useful to capture the normalized URL of the website and use this as a context feature. The normalized form was taken as the host name, with leading and trailing affixes removed. This feature was used with high confidence and no penalty for mismatch. Intuitively, it is highly likely that two people mentioned on the same site are the same person, though it is commonplace for a single person to be on multiple websites. Common URL's are given low information weight in the same way as any other feature in our system, so it is unnecessary to filter out uninformative sites.

3.2.2. Title words (stemmed/pseudo-stemmed/unstemmed)

Words in the title of a well constructed document often contain the necessary information for a human reader to disambiguate the focal entity of the document. These words were added as features. To help sparse information counts, these words were stemmed in three ways: Cpan's Lingua::Stem, the substring consisting of the initial four characters of all words longer than four characters, and left unstemmed. Best results were obtained using the substring method.

3.2.3. Lists

Lists of names occur in every type of text document that discusses named entities. There are various types of lists, such as author

lists, genealogies, teams, political districts, tour locations, organization members, disambiguation pages, et cetera.

If a name appears in the same, or similar, name list in two different documents, then there is a very high chance that the two names are co-referent. Likewise, a name that appears in a name list in one document has less chance of matching a name on its own in a different document.

The information given by a name list is conditional, not independent; correlation with a list is different from correlation with each item in the list separately. Using items separately causes redundancy. Using items together allows extra score for the collocation of the entities on the list. For example, if there are a set of five very common entities that rarely occur all together, it would improve the likelihood of match if they were treated as a single entity.

Many current entity disambiguation systems fail when they encounter documents which are entirely composed of lists, such as genealogies pages. An entity disambiguation system capable of correctly handling such documents would improve on the state of art performance level.

For this paper we used a simplistic method of detecting lists: look for sequences of entities with no string of alphabetic characters longer than three in between and then remove lists shorter than a specified number of items (in this case three). This set of lists was then further normalized by removing any list that did not occur in at least two separate documents.

3.3. Feature importance

In the discussion that follows, we attempt to assemble empirical evidence that backups up our claim that named entities are useful for disambiguating persons.

To test the importance of a given feature type in the disambiguation of a given type of entity, we measure the mutual information between entity type and feature type, which can be interpreted as the degree to which knowing the value of a feature of a particular type enables us to predict which entity we will see. Specifically, let E be the set of distinct standardized entity names in a corpus, and let F be all features of a particular type, we compute:

$$MI(E,F) \equiv \sum_{e \in E, f \in F} p(e,f) \log\left(\frac{p(e,f)}{p(e)p(f)}\right)$$

Here, e and f are standardized, undisambiguated feature values (e.g., "person entity John Smith" but not "person entity John Smith number 5"), under the simplifying assumption that disambiguation might change the scores, but not the ordering of feature types.

Table 1 Context Mutual Information

	Loc (4494)	Org (10658)	Per (9926)
Surrounding Words (30254)	1.33	2.77	2.77
Surrounding POS strings (7416)	1.45	3.29	3.27
Document Locations (4494)	1.92	3.05	3.13
Document Organizations (10658)	2.94	5.44	5.53
Document Persons (9926)	3.04	5.49	6.54
Document Topic (1023)	2.21	4.44	4.33

In Table 1, the columns are different types of entities to be disambiguated and the rows are context types. The parenthetical number in the headers is the number of distinct types of the given entity or context type, and the column/row intersection is the amount of mutual information that the context type has with standardized names of the entity type. For this example, the corpus is a collection of approximately fifty thousand newswire pages gathered from the Web and processed to identify article body and named entities. Surrounding words are the four words on either side of the given mention, stopping at other mentions as boundaries. Surrounding POS text are strings of part of speech tags from the three words on either side of a mention, also stopping at other mentions. Document Locations, Organizations, and Persons are all of the respective named entity in the same document. Document Topic is derived from a k-means clustering of corpus documents under a bag-of-words representation in which all named entity mentions have first been removed. The label associated with the cluster to which a document is assigned becomes its "topic" identifier. Note that there are a few artifacts from the method we used to compute this information. In real-world applications, there is some correlation between number of members of a class and the mutual information received, though mutual information is invariant to the dimensionality of a random variable's distribution. If more words were used, the resulting information might have been higher though the ranking of scores would probably remain similar. Also some information weight is gained by the ability to distinguish entities that in reality are unambiguous or trivial.

From the table, a number of our intuitions are supported though there are a few surprises as well. One of our intuitions that stands out in the table is the importance of Document Persons as a disambiguating feature. For all entity types, including even words, Document Persons are more useful than any other feature we tested for distinguishing between mentions. Another intuition that was supported was the low utility of Document Locations as disambiguating features. Locations span various entities so are not very useful to any. A surprise in the table was the importance of POS Strings. Although their importance for most entities is relatively low, it is still on the same order of magnitude as other useful information whereas we would have suspected little to no usefulness. We believe that some of this can be explained by the occurrence of certain frequent entities in stereotypical syntactic constructions, but the observed strength of this feature is nevertheless intriguing.

Finally, these results support our assertion that neighboring words are of relatively low utility for named entity disambiguation.

Surprisingly, words hold less disambiguating power for other words than do co-occurring person names or organizations.

4. PERFORMANCE

We submitted five configurations of our system to WePS2:

FICO_1: A run with an older version of our NER system
FICO_2: A run with URL, Lingua:Stem stemmed title tokens, and Lists
FICO_3: A run with URL, pseudo-stemmed title tokens, and Lists
FICO_4: A run with URL, and unstemmed title tokens
FICO_5: A run with URL, unstemmed title tokens, and Lists

run	BEP	BER	FMeasure_0.5_BEP-BER
ALL_IN_ONE_BASELINE	0.43	1	0.53
COMBINED_BASELINE	0.43	1	0.52
ONE_IN_ONE_BASELINE	1	0.24	0.34
FICO_1	0.76	0.69	0.66
FICO_2	0.83	0.62	0.68
FICO_3	0.85	0.62	0.7
FICO_4	0.84	0.63	0.68
FICO_5	0.84	0.62	0.68

Table 2: Results

Best results were obtained with our system using all of our new features. Future research will look further into the differences between these new features and what new features are useful for person disambiguation.

5. CONCLUSION

We have shown that our intuitions that contextual named entities are useful for disambiguating persons are correct. We have begun to examine the utility of new features such as URLs, title words, and lists of named entities and shown that they are indeed useful for disambiguating persons. With more advanced methods for feature extraction, we believe that these new features will have a greater impact on performance.

6. REFERENCES

[1] Agirre, E. and Soroa, A. (2007b). UBC-AS: A Graph Based Unsupervised System for Induction and Classification. (Paper presented at the Fourth International Workshop on Semantic Evaluations, SemEval-2007)

[2] Javier Artiles, Julio Gonzalo and Satoshi Sekine. 2009. WePS 2 Evaluation Campaign: overview of the Web People Search Clustering Task In 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, April.

[3] Artiles, J., Gonzalo, J. and Sekine, S. (2007). The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task. (Paper presented at Semeval 2007, Association for Computational Linguistics)

[4] Bagga, A. and Baldwin, B. (1998, August). Entity-based Crossdocument Coreferencing Using the Vector Space Model. (Paper presented at 17th International Conference on Computational Linguistics (CoLing-ACL), Montreal, Canada)

[5] Blume, M. (2005, May). Automatic Entity Disambiguation: Benefits to NER, Relation Extraction, Link Analysis, and Inference. (Paper presented at 1st International Conference on Intelligence Analysis, McLean, Virginia)

[6] Caid, W. and Oing, P. 1997. System and Method of Context Vector Generation and Retrieval. U.S. Patent No. 5,619,709.

[7] Gooi, C. H. and Allan, J. (2004, May). Cross-Document Coreference on a Large Scale Corpus. (Paper presented at Human Language Technology Conference (HLT-NAACL).Boston, Massachusetts.

[8] Paul Kalmar, Dayne Freitag, Matthias Blume. 2008. Finding Sense: A Comparison of Named Entity Disambiguation and Word Sense Discovery. In preparations.

[9] Paul Kalmar. 2008. Less is More: Advantages of Using Local Homogenous Data Sets in Natural Language Processing. Master's Thesis at San Diego State University.

[10] Kalmar, P. and Blume, M. (2007, June). FICO: Web Person Disambiguation Via Weighted Similarity of Entity Contexts. (Paper presented at the Fourth International Workshop on Semantic Evaluations, SemEval-2007)

[11] Knuth, D. E. 1998. The Art of Computer Programming, Volume 3: Sorting and Searching. (Addison-Wesley Professional)

[12] Mann, G. S. and Yarowsky, D. (2003, May). Unsupervised Personal Name Disambiguation. (Paper presented at Conference on Computational Natural Language Learning (CoNLL), Edmonton, Canada)

[13] Mitchell, A.; Strassel, S.; Przybocki, P.; Davis, J. K.; Doddington, G.; Grishman, R.; Meyers, A.; Brunstein, A.; Ferro, L. and Sundheim, B. (2004). Annotation Guidelines for Entity Detection and Tracking (EDT), Version 4.2.6. Retrieved February 28, 2008 from University of Pennsylvania, Linguistic Data Consortium Web site: http://www.ldc.upenn.edu/Projects/ACE/

[14] Neill, D.B. (2002). Fully automatic word sense induction by semantic clustering. Master's thesis, Cambridge University.

[15] Pantel, P. and Lin, D. (2002). Discovering word senses from text. (Paper presented at the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining)