FICO: Web Person Disambiguation Via Weighted Similarity of Entity Contexts

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

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

Abstract

Entity disambiguation resolves the many-to-many correspondence between mentions of entities in text and unique real-world entities. Fair Isaac's entity disambiguation uses language-independent entity context to agglomeratively resolve mentions with similar names to unique entities. This paper describes Fair Isaac's automatic entity disambiguation capability and assesses its performance on the SemEval 2007 Web People Search task.

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. Simultaneously, multiple identical mentions refer to distinct entities (Bagga and Baldwin, 1998).

Fair Isaac's 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 task (Artiles et al. 2007) 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 Fair Isaac's 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 the Fair Isaac 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 Fair Isaac's automatic entity disambiguation methodology and report on the performance of the system on the WePS data.

2 Methodology

In unstructured text, each document provides a natural context for entity disambiguation. After cleaning up extraneous markup we carry out within-document co-reference resolution, aggregating information about each entity mentioned in each document. We then use these entity attributes as features in determining which documents deal with the same entity.

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

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 149–152,
Prague, June 2007. ©2007 Association for Computational Linguistics

what text from those documents contains relevant information. As a first pass, the first HTML file in a folder which contained the query name was used as the main page. In retrospect, it might have been better to combine all portions of the page, or choose the longest page. We copied the title element and converted all text chunks to paragraphs, eliminating all other HTML and script. If no HTML was found in the directory for a page, the first text file which contained the query was used instead.

2.2 Within-Document Disambiguation

When dealing with unstructured text, a named entity recognition (NER) system provides the input to the entity disambiguation. Due to time constraints and that Persons are the entity type of primary interest, any mention that matches one of the query strings is automatically labeled as a Person, regardless of its actual type.

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 section will still merge some entities even within individual documents.

2.3 Cross-Document Disambiguation

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. In the domains of co-occurring locations, organizations, and persons, these 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. To reduce noise, only entities that occur within a given window of entities are included in this vector. In all test runs, this window is set to 10 entities on either side. Because of the effects that small corpora have on statistics, we added a large amount of newswire text to improve frequency counts. Many of the query names would have low frequency in a text corpus that is not about them specifically, but have high frequency in this task because each document contains at least one mention of them. This would cause the INF weight to incorrectly estimate the importance of any token; adding additional documents to the disambiguation run reduces this effect and brings frequency counts to more realistic levels.

We similarly count title tokens that occur with the entity and compute INF weights for the title tokens. Topic context, as described in Blume (2005), was used in some post-submission runs.

We define a separate distance measure per context domain. 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).

The process of cross-document entity disambiguation now boils down to repeatedly finding a pair of entities, comparing them (computing the sum of the above distance measures), and merging them if the score exceeds some threshold. We compute sets of keys based on lexical similarity and compare only entities that are likely to match. The WePS evaluation only deals with entities that match a query. Thus, we added a new step of key generation based on the query.

3 Performance

We have tested our entity disambiguation system on several semi-structured and unstructured text data sets. Here, we report the performance on the training data provided for the Web People Search task. This corpus consists of raw Web pages with substantial variation in capitalization, punctuation, grammar, and spelling – characteristics that make NER challenging. A few other issues also negatively impact our performance, including extraneous text, long lists of entities, and the issue of finding the correct document to parse.

The NER process identified a ratio of approximately 220 mentions per document across 3,359 documents. Within-document entity disambiguation reduced this to approximately 113 entities per document, which we refer to as *document-level entities*. Of these, 3,383 Persons (including those Organizations and Locations which were relabeled as Persons) contained a query name. Cross-document entity disambiguation reduced this to 976 distinct persons with 721 distinct standardized names. Thus, 2,407 merge operations were performed in this step. On average, there are 48 mentions per query name. Our system found an average of 14 unique entities per query name. In the gold standard, the average is 9 unique entities per query name.

Looking at the names that matched in the output, it is clear that NER is very important to the process. Post submission of our initial run, we used proper tokenization of punctuation and an additional NER system, which corrected many mistakes in the grouping of names. Also, many of the names that were incorrectly merged would not have been compared if not for the introduction of the additional key that compares all mentions that match a query name.

For the WePS evaluation submission, we converted our results to document-level entities by mapping each mention to the document that it was part of and removing duplicates. If we did not find a mention in a document, we labeled the document as a singleton entity.

We also used a number of standard metrics for our internal evaluation. Most of these operate on document-level entities rather than on documents. To convert the ground truth provided for the task to a form usable for these metrics, we assume that each entity contains all mentions in the corresponding document group. These metrics test the cross-document disambiguation rather than the NER and within-document disambiguation. These metrics should not be used to compare between different versions of NER and within-document disambiguation, since the ground truth used in the evaluation is generated by these processes.

In Table 1, we compare a run with the additional newswire data and the comparison key (our WePS submission), leaving out the additional newswire data and the additional comparison key, and leaving out only the additional comparison key.

In Table 2, we compare runs based on the improved NER (available only after the WePS submission deadline). The first uses the same parameters as our submission, the second uses an increased threshold, and the third utilizes the word vector-based clustering (document topics).

	Acc.	Prec.	Recall	Harm. Purity
WithExtraKey	0.670	0.545	0.906	0.818
NoAddedData	0.743	0.752	0.584	0.841
NoExtraKey	0.770	0.767	0.624	0.861

Table 1. Results of pairwise comparisons and clusterwise harmonic mean of purity and inverse purity on various disambiguation runs. Each metric is averaged across the individual results for every query name.

	Acc.	Prec.	Recall	Harm. Purity
WithExtraKey	0.690	0.618	0.552	0.815
1.25 Thresh	0.720	0.733	0.500	0.812
Topic Info	0.719	0.645	0.545	0.818

Table 2. Results based on improved named entity recognition. These should not be directly compared against those in Table 1, since the different NER yields different ground truth for these evaluation metrics.

Most of our metrics are based on pairwise comparisons – all document-level entities are compared against all other document-level entities that match the same query name, noting whether the pair was coreferent in the results and in the ground truth. With such comparison, we obtain measures including precision, recall, and accuracy. In this training data, depending on which NER is used, 35,000-50,000 pairwise comparisons are possible.

We also define a clusterwise measure of the harmonic mean between purity and inverse purity *with respect to mentions*. This is different from the metric provided by WePS, purity and inverse pu-

rity *at the document level*. Since some documents contain multiple entities, the latter metric does not perform correctly. Mentions, on the other hand, are always unique in our disambiguation. However, because the ground truth was specified at the document level, documents containing multiple entities that match a query yield ambiguous mentions. These decrease all purity-related scores equally and do not vary between runs.

The addition of the newswire data improved results. Inclusion of an extra comparison based on query name matches allowed for comparison of entities with names that do not match the format of person names, and only slightly reduced overall performance. The new NER run can only be compared on the last three runs. to the system performs better with topic context than without it.

In comparison, in the 2005 Knowledge Discovery and Dissemination (KD-D) Challenge Task ER-1a (the main entity disambiguation task), we achieved an accuracy of 94.5%. The margin of error in the evaluation was estimated at 3% due to errors in the "ground truth". This was a pure disambiguation task with no NER or name standardization required. The evaluation set contained 100 names, 9027 documents, and 583,152 pair-wise assertions.

4 Conclusions

Although the primary purposes of Fair Isaac's entity disambiguation system differ from the goal of the Web People Search task, we found that with little modification it was possible to fairly accurately cluster Web pages with a given query name according to the real-world entities mentioned on the page. Most of the errors that we encountered are related to information extraction from unstructured data as opposed to the cross-document entity disambiguation itself.

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 author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

References

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

Bagga, A. and Baldwin, B. (1998). Entity-based Cross-document Coreferencing Using the Vector Space Model. 17th International Conference on Computational Linguistics (CoLing-ACL). Montreal, Canada. 10-14 August, 1998, 79-85.

Blume, M. (2005). Automatic Entity Disambiguation: Benefits to NER, Relation Extraction, Link Analysis, and Inference. 1st International Conference on Intelligence Analysis. McLean, Virginia. 2-5 May, 2005.

Gooi, C. H. and Allan, J. (2004). Cross-Document Coreference on a Large Scale Corpus. Human Language Technology Conference (HLT-NAACL). Boston, Massachusetts. 2-7 May, 2004, 9-16.

Kalashnikov, D. V. and Mehrotra, S. (2005). A Probabilistic Model for Entity Disambiguation Using Relationships. SIAM International Conference on Data Mining (SDM). Newport Beach, California. 21-23 April, 2005.

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

Mann, G. S. and Yarowsky, D. (2003). Unsupervised Personal Name Disambiguation. Conference on Computational Natural Language Learning (CoNLL). Edmonton, Canada. 31 May - 1 June, 2003, 33-40.

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*. http://www.ldc.upenn.edu/Projects/ACE/.

Ravin, Y. and Kazi, Z. (1999). Is Hillary Rodham Clinton the President? Disambiguating Names across Documents. ACL 1999 Workshop on Coreference and Its Applications. College Park, Maryland. 22 June, 1999, 9-16.