Name-Ethnicity Classification from Open Sources

Anurag Ambekar
Dept. of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400
aambekar@cs.sunysb.edu

Charles Ward
Dept. of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400
charlesbward@gmail.com

Jahangir Mohammed
Dept. of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400
jahangir@cs.sunysb.edu

Swapna Male
Dept. of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400
maswapna@cs.sunysb.edu

Steven Skiena*
Dept. of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400
skiena@cs.sunysb.edu

ABSTRACT

The problem of ethnicity identification from names has a variety of important applications, including biomedical research, demographic studies, and marketing. Here we report on the development of an ethnicity classifier where all training data is extracted from public, non-confidential (and hence somewhat unreliable) sources. Our classifier uses hidden Markov models (HMMs) and decision trees to classify names into 13 cultural/ethnic groups with individual group accuracy comparable accuracy to earlier binary (e.g., Spanish/non-Spanish) classifiers. We have applied this classifier to over 20 million names from a large-scale news corpus, identifying interesting temporal and spatial trends on the representation of particular cultural/ethnic groups.

Categories and Subject Descriptors

I.2.1 [**Applications and Expert Systems**]: Cartography

General Terms

Algorithms, Experimentation

Keywords

ethnicity detection, name classification, news analysis, social science research

1. INTRODUCTION

Names are important, and can convey considerable information about people, places, and things. For example, we believe most readers have little difficulty proposing likely cultural/ethnic origins for people with names such as *Angelo Fasciano*, *Mohammad Al-Sabah*, *Luis Hernandez*, *Zvi Shamir*, *Dmitry Medvedev*, *Jiang Chen*, and *Akshay Ghandi*.

Here we describe our classifier for deducing the likely cultural/ethnic origins of arbitrary names. This work is part of our effort to develop the *Lydia* news/blog analysis system [14] into a resource for the social and political science research. Interactions between distinct ethnic/cultural groups comprise one of the dominant social forces shaping our world. Accurately quantifying the nature of these interactions provides essential data for social science research, in fields as diverse as history, political science, and international relations.

In this paper, we describe the methods underlying our *Lydia* name ethnicity classifier, as well as interesting analysis resulting from applying our classifier to large-scale news data. Table 1 demonstrates how our classifier deals appropriately with the names of twenty world leaders. We are by no means the first to build such a classifier; see our discussion of prior work in Section 2.

We do not claim our classifier is more accurate than others described in the literature because direct comparisons could not be conducted. However, informal comparisons suggest our 13 ethnicity classifier is quite competitive with simple binary (Spanish/non-Spanish) classifiers reported elsewhere. That said, we believe our classifier is interesting primarily because of its dependence on open data sources, methods, and because of our unique application to news analysis:

- *Dependence on Open Data Sources* – The most critical step in developing any ethnic name classifier is acquiring a large gold standard of ethnically-classified names. This is more difficult than it sounds. Racial/ethnic data is widely collected by government, industry, and other organizations, but sensitive personal data in such records are kept confidential. Significant genealogical and immigration records are typically made available only for small numbers of specific name queries, not large-scale downloading, and have issues with accuracy and significance.

  We have worked around these difficulties by basing it exclusively on data extracted from the open source encyclopedia Wikipedia [22]. Extracting reliable name-ethnicity pairs from Wikipedia required solving a variety of technical issues, but our methods generalize to

*Corresponding author. skiena@cs.sunysb.edu. Partially supported by NSF Grants EIA-0325123 and DBI-0444815.

KDD'09, June 28–July 1, 2009, Paris, France.

Name	Country	As GA GE	Ind GEA	Jap EA	Af Mus	Bri Jew WE EE	Nor Ita His Fre Ger
Cristina Fernández de Kirchner	Argentina	0.0 0.0 **1.0**				0.0 0.0 1.0 0.0	0.0 0.0 **.98** .02 0.0
Luiz Inácio Lula da Silva	Brazil	.01 .04 **.95**				.01 .01 **.93** .05	.02 .12 **.84** .02 .01
Giorgio Napolitano	Italy	0.0 0.0 **1.0**				0.0 0.0 **1.0** 0.0	0.0 **1.0** 0.0 0.0 0.0
Horst Köhler	Germany	0.0 0.0 **1.0**				0.0 0.0 **1.0** 0.0	0.0 0.0 0.0 0.0 **1.0**
Nicolas Sarkozy	France	0.0 0.0 **1.0**				0.0 .47 **.53** 0.0	0.0 0.0 0.0 **1.0** 0.0
Kevin Rudd	Australia	0.0 0.0 **1.0**				**.90** .10 0.0 0.0	
Stephen Harper	Canada	0.0 0.0 **1.0**				**.97** .02 0.0 0.0	
Gordon Brown	UK	0.0 0.0 **1.0**				**.98** .02 0.0 0.0	
Felipe Calderón	Mexico	0.0 0.0 **1.0**				**1.0** 0.0 0.0 0.0	
Dmitry Medvedev	Russia	0.0 0.0 **1.0**				0.0 0.0 0.0 **1.0**	
Viktor Yushchenko	Ukraine	.10 0.0 **.90**				0.0 0.0 0.0 **1.0**	
Shimon Peres	Israel	0.0 0.0 **1.0**				0.0 **.73** .22 .05	
Pratibha Patil	India	**1.0** 0.0 0.0	**1.0** 0.0				
Hu Jintai	China	**.99** .01 0.0	0.20 **0.8**	0.0 **1.0**			
Lee Myung-bak	South Korea	**1.0** 0.0 0.0	0.0 **1.0**	0.0 **1.0**			
Taro Aso	Japan	**.82** .15 .03	.02 **.98**	**.99** .01			
Barack Obama	USA	0.0 **1.0** 0.0			**1.0** 0.0		
Abdullah bin Abdul Aziz	Saudi Arabia	0.0 **1.0** 0.0			0.0 **1.0**		
Abdullah Gül	Turkey	.00 **.99** .01			0.0 **1.0**		
Asif Ali Zardari	Pakistan	0.0 **1.0** 0.0			0.0 **1.0**		

Table 1: World Leaders with the results of their ethnic classifications. Group abbreviations are described in Table 2.

classify named entities according to any of the hundreds of thousands of interesting groups defined in Wikipedia.

- *Improved Classification Methods* – Despite the enormity of Wikipedia, our extracted name-ethnicity pairs provide us with between one-to-two orders of magnitude fewer names than enjoyed by state-of-the-art classifiers. This mandates that we be more sophisticated in our use of training data.

  Our classifier is based on a hierarchical decision tree of hidden Markov models (HMMs). We make extensive use of k-mer frequency analysis to compensate for names not occurring within our corpus. We describe our techniques in detail, and present experiment results supporting a variety of design decisions.

- *Applications to News Analysis* – To demonstrate the significance of our ethnicity classification, we present several intriguing results in applying it to news streams, drawn from a large-scale study [23]. We uncover interesting trends in interactions between ethnic groups, as well as the frequency and sentiment associated with different groups in different contexts.

- *Research Availability of our Classifier* – We did not set out initially to develop our own classifier, but our inability to gain sufficient access to commercial or even research classifiers forced our hand. In response, we are committed to maintain our ethnicity classification server (*http://www.textmap.org/ethnicity*) as an open resource for all legitimate research applications.

Our paper is organized as follows. We discuss related work in name ethnicity analysis in Section 2. Methods for extracting reference data from Wikipedia are discussed in Section 3. The HMM and decision tree techniques underlying our classifier are presented in Section 4. Performance results on the accuracy of our classifier and the impact of various design decisions is reported in Section 5. Interesting applications of our analysis to large-scale news feeds follow in Section 6. Finally, our conclusions and future work are presented in Section 7.

2. PREVIOUS WORK

The problem of ethnicity identification from names has a variety of significant applications. Perhaps the most important application is to biomedical research [6, 4, 7, 8, 20]. The notion of race or ethnicity can serve as a surrogate reflecting the genetic or dietary differences among individuals. Experimental design criteria may block direct requests for ethnic-self identification, or the need for such data may not be realized until too late in the course of a study. Other applications include population demographic studies [12, 16], including predicting overall identity status [1] and measuring the eradication of discrimination [17]. Finally, there are business applications in targeting product marketing efforts towards particular groups of individuals. Mateos [16] provides an excellent review of work in the ethnicity detection.

An important problem faced by any such classification program is identifying a meaningful taxonomy of linguistically-distinct cultural/ethnic groups, a task requiring extensive knowledge of history, linguistics, and demographics. Here we have been most impressed by the work of Mateos, Webber, and Longley [17], who propose a taxonomy into Cultural, Ethnic and Linguistic (CEL) types. Their taxonomy consists of 185 CEL types which are grouped into 15 broad CEL groups. These groups correspond closely to the ethnic groupings recognized by our classifier.

Many previous efforts in ethnicity classification focuses on distinguishing members of a single CEL group. Binary classifiers can be very useful to particular domains. The *Generally Useful Ethnicity Search System* (GUESS) uses Spanish names to determine Hispanic ethnicity [5]. Coldman, Braun, and Gallagher perform binary classification of names as either Chinese or non-Chinese [6]. SANGRA (*South Asian Names and Group Recognition Algorithm*) is one of several programs to identify South Asians by names, which are analyzed in [10, 11]. However, for the purposes of our analysis, we require a multi-faceted classification system, capable of discriminating among a relatively large number of ethnic groups. As noted by Mateos [16], this type of classification engine has seen relatively little attention.

The primary method used in ethnicity classification is comparison to known surname lists. Deriving accurate

name lists for use in a classification engine requires substantial efforts, however. Lauderdale and Kestenbaum describe the development of surname lists for six Asian ethnicity groups [12]. Coldman et al. [6] use a simple probabilistic method based on full name lists which is somewhat similar to our system. This system could not, however, classify names which did not appear in the input name lists. Fiscella and Fremon [7] describe the combination of geocodal information and surname analysis to predict the ethnicity. They state that better results are achieved by combining both the geocodal and surname information. In subsequent sections, we will describe methods for the compilation of our name lists, the classification methods which utilize them, and the accuracy of our resulting system.

3. EXTRACTING NAME LISTS FROM WIKIPEDIA

Wikipedia is an open repository containing much of the world's general knowledge, which makes it an attractive focus for projects on information extraction. In this case, we needed to extract fairly accurate lists of names labeled by ethnicity, but more generally we would like to produce lists of entities corresponding to members of any natural group [19].

We sought to exploit Wikipedia's extensive network of categories under which articles are classified. For example, basketball player "Steve Nash" is classified as member of the following categories:

> 1974 births | Living people | Basketball players at the 2000 Summer Olympics | British Columbia sportspeople | Canada's Walk of Fame | Canadian basketball players | Canadian expatriate sportspeople in the United States | Point guards | Dallas Mavericks players | Canadians of British descent | Members of the Order of British Columbia | Officers of the Order of Canada | Olympic basketball players of Canada | People from Victoria, British Columbia | People from Johannesburg | Canadians of Welsh descent | Phoenix Suns players | Santa Clara Broncos men's basketball players | South African immigrants to Canada | Santa Clara University alumni.

From these categories, we might hope to learn that he is a basketball player and possesses a name of Welsh ancestry. At the time of this study, Wikipedia's roughly three million articles were classified into exactly 293,119 categories with 610,032 edges/relationships defined between them, making this a valuable potential source of ontologies.

Our task is complicated by several factors. First, the category names are not systematically constructed, so it is difficult to identify the root category associated with a particular group. Second, membership in interesting categories (say "basketball player") is often not explicitly defined but implied by category inclusion. Third, the network of categories is a directed graph with cycles, not a hierarchy or a DAG, rendering it ill-defined to expand categories by inclusion. Finally, we note that irregularities result because subcategories are not explicitly defined by "is a" relationships. For example, "Serial Killer Movies" is a sub-child of the category "Serial Killers", which should be semantically restricted to people.

We employed the following procedure to build our reference lists:

- *Network extraction* – We used WikiPrep (available at *http://sourceforge.net/projects/wikiprep*) to process the raw Wikipedia file dumps. WikiPrep is a set of scripts which extracts various content, including a network of Wikipedia categories. Parsers and other scripts were written to extract information in the required format.

- *DAG-ification and category expansion* – The category network from Wikipedia contained directed cycles. To break cycles, we added a root r to all source nodes (those without incoming edges) and performed a topological sort from r, deleting any back edges encountered. Only 756 back edges were identified, essentially all of which corresponded to non-"is a" relationships.

  With the graph now a DAG, we perform a DFS/BFS from any target category node to identify all child categories. The leaf nodes under these categories corresponded to Wikipedia article titles. The set of leaf nodes under the target categories were collected under the target category label.

- *Relevant Category Identification* – The next task was formulating a small list of broad categories to cover most interesting entities. Here we exploit our roster of categorized named entities extracted by *Lydia* from an extensive corpus of news documents.

  To select interesting categories, we used heuristic criteria such as the number of articles or named *Lydia* entities below each category, the number of words in the category title, and elimination of mechanical subclasses of categories refined by date or place (e.g. '1974 Births' and 'British Columbia sportspeople'). We also used external ontologies such as the most popular categories on Wikipedia, and Yahoo/Google's web directories to refine our lists. All these measures, and some manual editing, were used to create a hierarchically organized category set.

- *Entity Prunning* – Expanding groups across edges which do not correspond to "is a" relationships lead to inconsistencies, as in the "Serial killers" example give above. By using entity-type data derived from *Lydia* analysis of news text, we could identify and automatically eliminate edges which linked groups containing entities of different types (e.g. Person vs. Title), thus keeping the expanded groups relatively homogeneous.

- *Post Processing* - Wikipedia names are by no means perfectly characterized for our task. Consider the case of a cricket player of Indian origin who resides in England. We would interpret Wikipedia as classifying this name as being of English origin. Similar problems are discussed in [8]. Thus, a certain amount of the inaccuracy of our classifier should be attributed to such mislabeled data. To reduce this problem we manually cleaned some of the data-sets, reviewing certain ethnicities and removing any obvious misassignments. Representative of these problems were the prevalence of Muslim names within the Jewish name roster, because of the significant Arab population in Israel.

Table 2 gives the number of names in each ethnic groups used for training and description of corresponding Wiki-categories.

Ethnic Group	Abbrv.	Number of Names raw	clean	curated	National Seed Wiki-Categories
African	Af	4807	4304	3819	African Nations (Algerians, Nigerians, etc.)
British	Bri	43581	43581	39735	English, Welsh, Scottish, Irish, African Americans
EastAsian	EA	6526	6526	5849	Chinese, Koreans
EastEuropean	EE	9318	9318	8390	Russians, Belarusians, Bulgarians, Czechs
French	Fre	14076	14076	11274	French
German	Ger	4661	4632	3617	Germans
Hispanic	His	11365	11365	9717	Portuguese, Spaniards
IndianSubContinent	Ind	9356	9050	8096	Indians, Sri Lankans
Italian	Ita	13633	13633	11821	Italians
Japanese	Jap	8425	8425	7815	Japanese
Jewish	Jew	11332	11225	8885	Israelis, British Jews
Muslim	Mus	7069	6854	6449	Arabic Names
Nordic	Nor	5242	5242	4670	Swedes, Norwegians, Danes

Table 2: Training data extracted from Wikipedia, presenting the name count per ethnic group.

4. ALGORITHMIC METHODS

Our classifier consists of two major components: (1) a set of non-deterministic automata which map name strings to ethnic categorizations, and (2) a decision tree reflecting the historical/linguistic origins and coarse groupings of each ethnic group (see Figure 1). Starting at the root, the automaton is applied at intermediate nodes in the tree to decide which child to branch to. The identity of the leaf node on this root-to-leaf path through the decision tree defines the classified ethnicity of the given name.

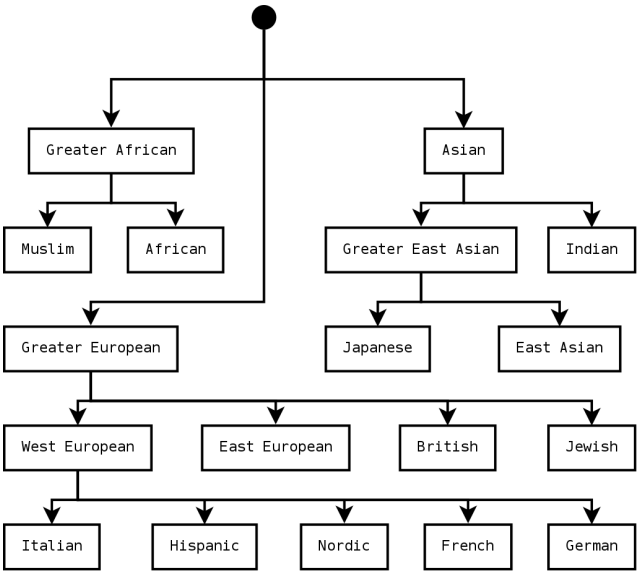

Figure 1: Hierarchical categorization of CEL groups. We classify names at each level of the tree, resolving identity at the leaves.

4.1 Hidden Markov Model

Our classification engine consists of a series of sub-classifiers operating at each level of the CEL group hierarchy. Each classifier is a Hidden Markov Model (HMM) automaton, with states corresponding to the position within a name, and observations corresponding to substrings from the name. We divide each full name into three parts: first-name, middle-name and last-name. When only two words exists in the full name, we consider only the first-name and last-name. When only one word exists, we declare it to be the last name.

For a classifier discriminating between e ethnicities, the states of the classifier constitute a directed acyclic graph with e subgraphs, each corresponding to a single ethnicity. Each of these paths has edges with observations consisting of strings of letters determined by the strings' relative frequency within the training data for that ethnic group.

Observations along edges consist both of full name-parts (e.g., "John") and substrings thereof (e.g., "Jo" or "hn"). The set of these observations, as well as the transition probabilities along these edges, determine the likelihood of a particular name being generated by the portion of the automaton corresponding to each ethnicity. Transition probabilities along multiple possible paths were manually determined, and biased towards paths which consume larger numbers of letters. That is, the system will prefer paths which correspond to names which appear in the input data over those which it assembles from prefixes and suffixes, all other things being equal. Figure 2 shows diagrammatically the construction of an individual automaton.

The dark-gray states in Figure 2 correspond to matching a name-part which occurs in the training data set. The set of observations and observation probabilities on the transitions to this state are made up of full name-parts and their associated frequency within this ethnicity in the training dataset. The initial white state corresponds to the case in which the full name-part does not occur within this ethnicity in the training dataset. The transition probabilities to these two states are biased such that there is a large preference for paths which consume full names.

The remaining states correspond to matching k-mer components (k length substrings) of the name-part. Here, entering each light-gray state corresponds to matching an individual k-mer, while entering each white state corresponds to a failure to match a k-mer. Observations and observation probabilities on each light-gray state correspond to the relative frequency of each k-mer within the training data for the ethnicity. Observation probabilities on each white state are evenly split among all possible k-mer observations. Transition probabilities were chosen to bias heavily towards

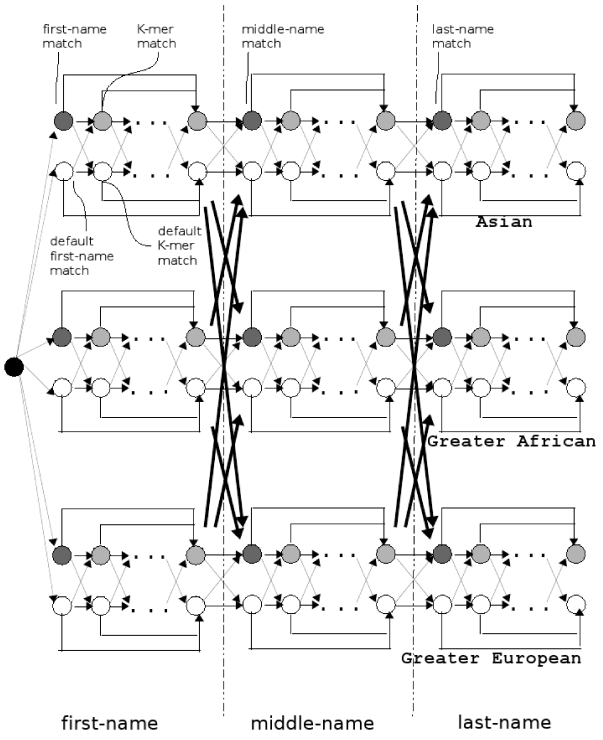

Figure 2: Top-level automaton of the ethnicity classifier. Its three stages match the first-name, middle-name and last-name respectively for the *Asian*, *Greater African* and *Greater European* CEL groups.

matching each k-mer through a light-gray state, if possible, and thusly penalize the use of the non-matching white states.

The observation probability $f(N)_E$ for the full name-part N match (dark-gray state) in ethnicity E is calculated as:

$$B = f(N)_E / (\sum_{E'} f(N)_{E'}) \quad (1)$$

The observation probability for the k-mer match s (light-gray state) in ethnicity E is calculated similarly. The probability distribution on each of these k-mers is determined not only by the distribution of k-mers in the ethnic names, but also on the k-mers' positions within these names. That is, the appearance of certain k-mers are strongly biased towards the beginning or end of names, and this is reflected in the observation probabilities of the k-mer match states.

The automaton is also constructed to allow a path which "blends" ethnicities. That is, the given name and the surname of a name may indicate two separate ethnicities. The automaton allows this by allowing a path to move between ethnicities between first/middle/last name boundaries, at some cost. Parameter settings allow us to vary the relative significance of the first, middle, and last names in the final analysis. A particular name is assigned the label of the ethnic group associated with the final state with the largest probability.

4.1.1 Prefix/Suffix k-mers

A particular significant set of k-mers for ethnicity classification are associated with the prefix/suffix strings of both family and given names. The prefix "Mac" suggests Scottish or Irish ancestry (British CEL group, in our categorization) far more strongly than the 3-mer in general. Moreover, as our system uses 2-mers in its optimized configuration, this could be missed entirely.

String	Type	Ethnicity	Freq	Score S
Al-	prefix	Muslim	391	0.96
Singh	full	IndianSubContinent	225	0.93
Yoshi	prefix	Japanese	165	0.96
Yama	prefix	Japanese	145	0.91
Naka	prefix	Japanese	140	0.94
Bha	prefix	IndianSubContinent	132	0.94
mura	suffix	Japanese	178	0.97
kawa	suffix	Japanese	147	0.99
ensen	suffix	Nordic	102	0.98
uan	suffix	EastAsian	102	0.95
ikov	suffix	EastEuropean	102	0.94

Table 3: Most significant prefix/suffix rules for name ethnicity classification.

To exploit this valuable cue, we compute prefix/suffix probabilities independently from other k-mers to define the appropriate transitions in the classifier. In particular, we sought a set of prefixes and suffixes providing some certainty that the names belong to a particular ethnicity. Table 3 lists some of the most significant prefixes/suffixes we identified in this search. We experimented with prefix/suffix lengths ranging from three to five, and identified the frequency $f(p)_E$ as the number occurrences of prefix/suffix p within distinct names of the given ethnicity E. The significance score S of this frequency is analogous to the observation probabilities on states as given in Equation 1.

4.2 Decision Tree

We broadly follow the basic CEL groups identified by [17], though we depart in several minor ways so as to allow for a more natural hierarchical categorization of these groups. The reason is three-fold. First, this hierarchical categorization allowed us to analyze the accuracy of our classification engine at multiple levels of granularity as well as identify name groups which were mutually indistinct. Second, efficiency issues arise in dealing with classification over our 20 million name news corpus which favor batch processing of smaller sets of groups. Finally, we must compensate for the sparsity of data about certain groups. The hierarchical classification provides us with reasonable aggregation levels for this purpose, allowing us to vary the trade-off between granularity and accuracy in our analysis.

Figure 1 shows our hierarchical categorization of CEL groups. Among the changes with respect to the Cultural Ethnic and Linguistic (CEL) groups of [17], we have collapsed *Spanish* and *Portuguese* together to form a *Hispanic* group. We replaced the *English* group with a broader *British* group including *Celtic*, *English* and *African American*. Finally, we have placed the *Jewish* group within the broad group *European*.

5. PERFORMANCE RESULTS

To assess the accuracy of our classification engine, we constructed automatons for each level of classification us-

ing training sets comprising 70% of the data extracted from Wikipedia and tested against the remaining 30%. Figure 3 shows the precision, recall, and F-scores for each ethnicity as computed by the hierarchical classifier.

In general, the accuracy at the coarsest level of granularity is quite good, while at the finest it is comparable to other published ethnic classification systems. For example, the performance of our system in classifying Hispanic names is roughly equivalent to the accuracy of GUESS (as assessed by Stewart et al. [21]). The most significant practical failing of the classifier is in the classification of Jews. The classifier here has quite low accuracy, and is heavily confused with the British CEL group, resulting in the Jewish group being artificially inflated. Table 5 gives the total entity counts and breakdown by total names and total entity references of each ethnicity as determined by our classifier on our U.S. daily newspaper corpus.

We find that we did particularly well in ethnic groups like *British*, *Japanese*, *IndianSubContinent*, *EastAsian* (*Chinese*, *Korean*). This is because these names are fairly distinct and our classifier identifies this difference. In cases such as *Jewish* or *Africans*, our classifier is not quite as good, because of forces of history and colonization blur the interpretation of name as a surrogate for ethnicity. Additionally, small training sets from the wikipedia data for the groups, such as the *German* and *Jewish* groups, also worsened performance for these groups.

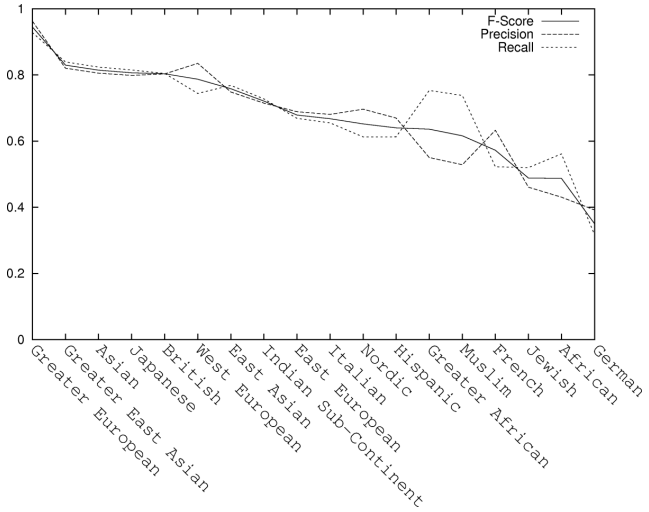

Figure 3: Precision, recall and F-score curves for all the CEL groups in the hierarchy, with ethnicities sorted by F-score.

5.1 Experimental Variations

To optimize the performance of our classifier, we experimented with five design decisions in the construction of the classifier and evaluated their impact on precision/recall:

- *Best Path vs. Sum of Paths* – The CEL group can be assigned according to the terminating state with highest probability; in other words, the state with the greatest weighted sum of all paths ending in it. Alternatively, the CEL group can be assigned according to to the terminating state which has the single most probable path. The best path method gives somewhat better accuracy than sum of paths.

- *Single-Tier vs. Hierarchical Classifier* – In addition to the hierarchical classifier, we also experimented with a single-tier classifier which attempts to discriminate between all 13 leaf ethnicities simultaneously. We find that for some ethnicities the single-tier works better, while for others the hierarchical classifier is best. The results favor single-tier for *EastAsian*, *German*, and *French* while the tree-based classifier gives better results for *British*, *Muslim*, and *Jewish*. However, we lose the facility of getting broader classification from internal nodes with a single-tier classifier. Moreover, with the single-tier architecture, the large number of states slows the processing through the automata significantly.

- *Equal vs. Biased-Priors* – Our initial classifier used equal transition probabilities to move to each initial state. We also experimented with transitions probabilities determined by relative frequency of names in each ethnicity. Having known estimates of population density has proved useful for binary classifiers [6]. In general, this method increased the precision of groups where we had small training sets (*German*, *Muslim*, and *African*), at the expense of recall.

- *Aggressive vs. Passive Curation* – This experiment attempted to improve a classifier by using it to clean its training data. We tested each name from the initial training data-set using the initial classifier, and then removed all the wrongly-classified names from the training data-set. We then re-built the classifier using this *curated* data. Table 2 gives the number of names in the original training set and in the *curated* training set. This procedure proved disappointing, however, performing poorly compared to baseline; it particularly hurt the classification of ethnicities with small training sets.

- *2-mers vs. 3-mers* – The final experiment we performed concerned the difference between various values of k in the construction of our k-mers. We find that increasing k decreased the overall accuracy of the classifier.

Overall, the single-tier best-path classifier arguably performed best. This is not, however, the final classifier used in our news analysis. This is because, as previously mentioned, the hierarchical classifier has several practical advantages we wished to exploit. Thus, we use the hierarchical best-path classifier for our news analysis.

6. APPLICATIONS TO NEWS STREAMS

Our work on ethnic name classification was motivated by our goal to develop the *Lydia* news/blog analysis system as a resource for research in the social sciences. *Lydia* [2, 3, 9, 14, 13, 15, 18] employs natural language processing (NLP) and statistical analysis to reduce text streams to time series data on the news volume and sentiment associated with each news entity. See *www.textmap.org* for a representation of our analysis.

Ethnic Group	Hierarchical Best-Path			Single-Tier Best-Path			Sum of	Biased	Curated	k-mer
	Precision	Recall	F-Score	Precision	Recall	F-Score	Paths	Priors	Data	k=3
Greater European	0.96	0.93	0.95	N/A	N/A	N/A	−0.3%	+0.4%	−0.1%	+0.2%
Greater African	0.56	0.75	0.65	N/A	N/A	N/A	−1.4%	+2.7%	−0.9%	−2.4%
Asian	0.82	0.83	0.82	N/A	N/A	N/A	−1.1%	+0.5%	−0.1%	+0.6%
Greater East Asian	0.84	0.84	0.84	N/A	N/A	N/A	−1.1%	+0.4%	+0.1%	+0.7%
Western European	0.84	0.75	0.79	N/A	N/A	N/A	−0.4%	+0.5%	−0.8%	−1.0%
African	0.44	0.56	0.49	0.48	0.59	0.53	−0.9%	+3.1%	−1.4%	−2.6%
British	0.80	0.81	0.81	0.82	0.77	0.79	−0.4%	+0.1%	−0.5%	−0.3%
East Asian	0.77	0.77	0.77	0.75	0.79	0.77	−1.8%	+0.4%	+0.2%	+1.5%
EastEuropean	0.70	0.69	0.69	0.72	0.71	0.72	−12.0%	+1.0%	−0.8%	−2.1%
French	0.64	0.53	0.58	0.61	0.54	0.57	−1.2%	+0.1%	−1.8%	−7.0%
German	0.40	0.33	0.36	0.34	0.49	0.41	−3.6%	+3.9%	−4.7%	−7.5%
Hispanic	0.68	0.62	0.65	0.63	0.49	0.63	−1.2%	−0.3%	−1.2%	−4.6%
Indian Sub-Continent	0.73	0.73	0.73	0.72	0.74	0.73	−1.6%	+1.4%	−0.6%	+0.6%
Italian	0.69	0.66	0.67	0.65	0.68	0.66	−0.9%	+0.0%	−0.6%	−3.6%
Japanese	0.81	0.82	0.82	0.83	0.82	0.83	−1.1%	+0.4%	+0.4%	+1.4%
Jewish	0.46	0.52	0.49	0.52	0.49	0.50	−0.5%	+1.4%	−1.5%	−0.2%
Muslim	0.55	0.75	0.63	0.62	0.70	0.66	−2.4%	+3.6%	−1.1%	−2.0%
Nordic	0.71	0.63	0.67	0.58	0.66	0.62	−2.3%	−0.1%	−1.8%	−3.8%

Table 4: Precision, Recall, and F-Score of the two best classifiers, and % change in F-Score for various experiments with the hierarchical classifier

Ward, et.al. [23] have used our classifier in the course of a significant study on differences in news coverage between cultural groups. We present a small portion of their data here to provide the flavor of this analysis and illustrate the significance and utility of our name classifier. Table 5 captures the scale of this study showing the classifications of over 20 million distinct names, comprising over 100 million individual entity references in our text corpus.

CEL Group	Entity Count	% of Names	% of Refs
Greater European	19,590,638	88.3%	90.4%
Greater African	1,013,192	4.6%	4.4%
Asian	1,575,992	7.1%	5.2%
Greater East Asian	863,058	3.9%	3.1%
West European	7,450,054	33.9%	30.8%
African	498,185	2.2%	2.0%
British	11,354,975	51.1%	56.7%
East Asian	508,293	2.3%	1.9%
East Euro.	785,609	3.5%	2.9%
French	1,105,933	5.0%	3.8%
German	555,810	2.5%	2.0%
Hispanic	1,160,509	5.2%	5.0%
Italian	1,252,638	5.6%	5.0%
Indian	712,934	3.2%	2.1%
Japanese	354,765	1.6%	1.2%
Jewish	2,954,538	13.3%	13.4%
Muslim	515,007	2.3%	2.4%
Nordic	420,626	1.8%	1.6%

Table 5: Distribution of entities by classified ethnicity as % of distinct person entities and % of news coverage

Each name was subjected to ethnic classification using the program described in this paper, and the reference/sentiment counts aggregated over all entities within each CEL group. Figure 4 presents time series showing the relative frequency and sentiment of CEL groups over the past four years as represented within U.S. daily newspapers. Interesting phenomena include large seasonal fluctuations in the number of references to Hispanics (attributable to their representation in professional baseball) and persistently lower sentiment associated with Muslims and (to a lesser extent) Hispanics than members of the British and French CEL groups.

By associating each news source with its point of origin and modeling its sphere of influence, we can identify spatial trends in reference and sentiment using the methods of [18]. Figure 5 demonstrates the uncanny agreement between our modeling of the frequency of Hispanic names (as identified using our classifier) in newspapers and U.S. Census data on the distribution of Hispanics within the United States. Results like these confirm both the general accuracy of our name classifier and its utility for social science research.

7. CONCLUSIONS AND FUTURE WORK

We have presented a methodology for name-based ethnicity classification from open sources, with experimental results governing performance. In general, we are pleased with our results, as our classifier performs fairly well except for some ethnicities where the training data is limited or ambiguous.

One direction for further work involves improving the precision of our classifier as well resolving finer distinctions between ethnic groups (e.g. distinguishing between Russian and other Eastern European groups). The key issue is obtaining a more precise and comprehensive training data-set. One approach we have been evaluating is using person/place collocations in our extensive news corpus to name each name to nationality, thus generating a significantly larger (albeit noisier) training set.

Our other main focus is in applying our classifier more extensively to news data to analyze historical trends in in-

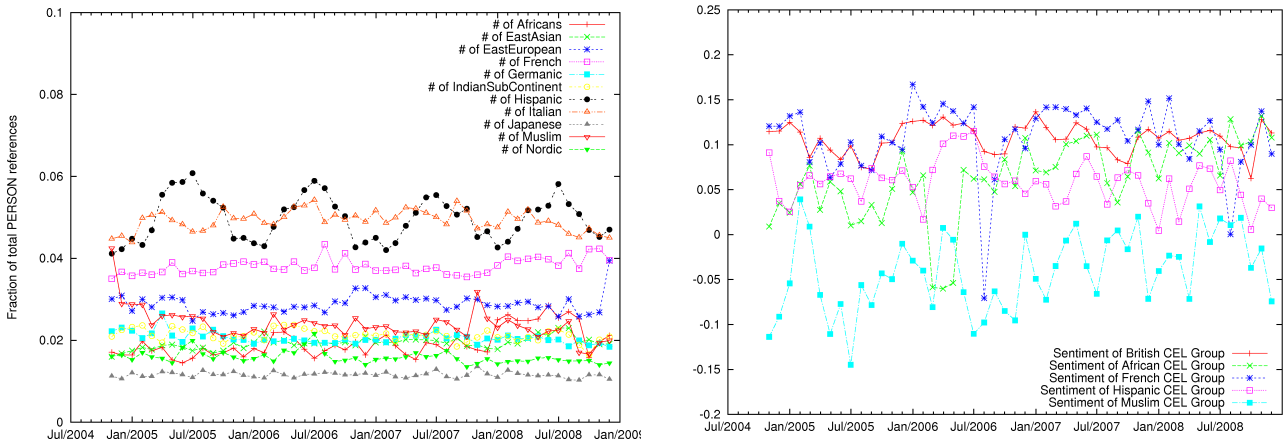

Figure 4: Frequency and sentiment time series data for CEL groups over four years of U.S. daily newspapers.

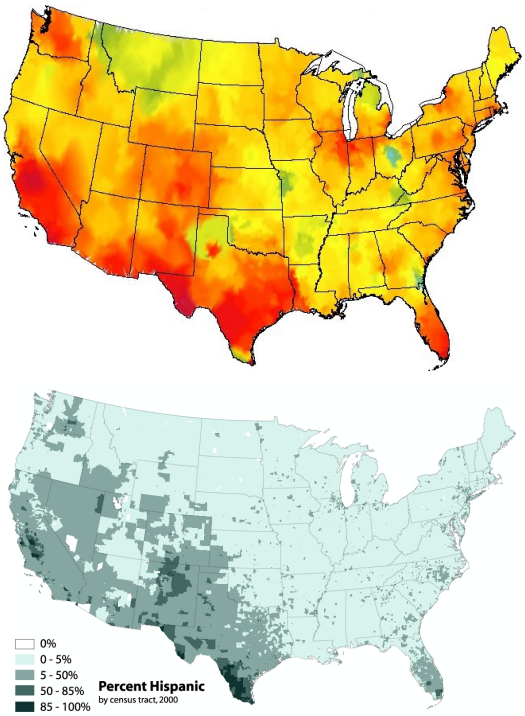

Figure 5: Comparison of the distribution of Hispanics in the United States according to news analysis (top) and the U.S. Census (bottom).

ternational relations, intergroup relations, and racial/ethnic discrimination. We anticipate that this analysis will be of great interest to a wide class of social scientists.

8. REFERENCES

[1] E. Aries and K. Moorehead. The importance of ethnicity in the development of identity of black adolescents. *Psychological Reports*, 65:75–82, August 1989.

[2] M. Bautin and S. Skiena. Concordance-based entity-oriented search. In *IEEE/WIC/ACM Int. Conf. Web Intelligence (WI-07)*, pages 586–592, 2007.

[3] M. Bautin, L. Vijayarenu, and S. Skiena. International sentiment analysis for news and blogs, 2008.

[4] E. Berchard, E. Ziv, and et. al. Importance of race and ethnic background in biomedical research and clinical practice. *The New England Journal of Medicine*, 348:1170–1175, March 2003.

[5] R. W. Buechley. Generally useful ethnic search system, GUESS. In *Presented at the Annual Meeting of the American Names Society*, New York, NY, 1976.

[6] A. J. Coldman, T. Braun, and R. P. Gallagher. The classification of ethnic status using name information. *Journal of Epidemiology and Community Health*, 42:390–395, 1988.

[7] K. Fiscella and A. M. Fremon. Use of geocoding and surname analysis to estimate race and ethnicity. *Health Service Research*, 41:1482:1500, August 2006.

[8] P. Gill, R. Bhopal, S. Wild, and J. Kai. Limitations and potential of country of birth as proxy for ethnic group. *British Medical Journal*, 330:196, 2005.

[9] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-Scale Sentiment Analysis for News and Blogs. In *Proc. First Int. Conf. on Weblogs and Social Media*, pages 219–222, Mar. 2007.

[10] S. Harding, H. Dews, and S. Simpson. The potential to identify South Asians using a computerised algorithm to classify names. *Population Trends*, 97:46–9, 1999.

[11] D. Honer. Identifying ethnicity: A comparison of two computer programmes designed to identify names of south asian ethnic origin. *MPH Dissertation, University of Birmingham*, 2003.

[12] D. S. Lauderdale and B. Kestenbaum. Asian american ethnic identification by surname. *Population Research and Policy Review*, 19:283–300, 2000.

[13] L. Lloyd, P. Kaulgud, and S. Skiena. Newspapers vs. blogs: Who gets the scoop? In *Computational Approaches to Analyzing Weblogs (AAAI-CAAW 2006)*, volume AAAI Press, Technical Report SS-06-03, pages 117–124, 2006.

[14] L. Lloyd, D. Kechagias, and S. Skiena. Lydia: A system for large-scale news analysis. In *String Processing and Information Retrieval (SPIRE 2005)*, pages 161–166, 2005.

[15] L. Lloyd, A. Mehler, and S. Skiena. Identifying co-referential names across large corpra. In *Proc. Combinatorial Pattern Matching (CPM 2006)*, volume LNCS 4009, pages 12–23, 2006.

[16] P. Mateos. A review of name-based ethnicity classification methods and their potential in population studies. *Population, Space and Place*, 2007.

[17] P. Mateos, R. Webber, and P. Longley. The cultural, ethnic and linguistic classification of populations and neighbourhoods using personal names. Technical report, CASA Working Papers 116, Centre for Advanced Spatial Analysis University College London, March 2007.

[18] A. Mehler, Y. Bao, X. Li, Y. Wang, and S. Skiena. Spatial Analysis of News Sources. In *IEEE Trans. Vis. Comput. Graph.*, volume 12, pages 765–772, 2006.

[19] A. Mehler and S. Skiena. Expanding network communities from representative examples. *ACM Trans. Knowledge Discovery from Data (TKDD)*, 2009.

[20] K. Nanchahal, P. Mangtani, M. Alston, and I. dos Santos Silva. Development and validation of a computerized South Asian names and group recognition algorithm (SANGRA) for use in british health-related studies. *Journal of Public Health Medicine*, 23:278–285, 2001.

[21] S. L. Stewart, K. C. Swallen, S. L. Glaser, P. L. Horn-Ross, and D. W. West. Comparison of Methods for Classifying Hispanic Ethnicity in a Population-based Cancer Registry. *Am. J. Epidemiol.*, 149(11):1063–1071, 1999.

[22] J. Wales. Wikipedia. http://www.wikipedia.org, 2009.

[23] C. Ward, M. Bautin, and S. Skiena. Identifying differences in news coverage between cultural/ethnic groups. submitted for publication, 2009.