A Data-driven Approach to Studying Given Names and their Gender and Ethnicity Associations

Shervin Malmasi
Centre for Language Technology
Macquarie University
Sydney, NSW, Australia
shervin.malmasi@mq.edu.au

Mark Dras
Centre for Language Technology
Macquarie University
Sydney, NSW, Australia
mark.dras@mq.edu.au

Abstract

Studying the structure of given names and how they associate with gender and ethnicity is an interesting research topic that has recently found practical uses in various areas. Given the paucity of annotated name data, we develop and make available a new dataset containing 14k given names. Using this dataset, we take a data-driven approach to this task and achieve up to 90% accuracy for classifying the gender of unseen names. For ethnicity identification, our system achieves 83% accuracy. We also experiment with a feature analysis method for exploring the most informative features for this task.

1 Introduction

The study of the structure and phonology of given names and how they associate with gender and ethnicity is a topic that has been investigated over the last several decades (Cassidy et al., 1999). Such research generally aims to identify discriminative features that are predictive of gender and how they can be applied in other areas.

In addition to linguistic and psychological research, gender discrimination and inference has recently found practical uses in various areas. These include applications in NLP and text mining, as outlined in §2.

These applications generally rely on name databases which can be costly to construct. However, maintaining an exhaustive database is not always feasible for a number of reasons. Exact matching may not be possible due to spelling variations, portmanteaus and self-created nicknames. Names found on the Web can also be combined or shortened, particularly in usernames.

The main objective of this work is to evaluate a data-driven approach to inferring gender and ethnicity using machine learning and probabilistic models. Such methodology can help inform researchers about the structure of names and to build more robust name-gender inference systems for classifying unseen names.

In line with these goals, we also develop a publicly available dataset of names annotated with gender and ethnicity information.

2 Related Work

Linguistic research in name-gender relations has yielded some interesting results and demonstrated that phonology is quite informative about the gender of a name (Cassidy et al., 1999). It has also suggested that humans can often make gender attribution errors about other people due to the phonetic characteristics of their names.

In fact, through manual analysis, researchers have identified a slew of features strongly associated with gender (Barry Jr and Harper, 1995; Cutler et al., 1990), often from a linguistic viewpoint. One such example is that English female names are far more likely to end in a schwa vowel than male names (Slater and Feinman, 1985), most likely due to their Latin etymology. It is believed that humans implicitly learn these phonological cues and gender associations through exposure.

Names are also a useful demographic indicator to identify people that fit certain criteria or are members of particular groups. For example, manual analysis is often applied to phone directory or other data to identify potential candidates from a specific background for a biomedical study (Yavari et al., 2006).

Such applications have also expanded to language technology, given that names are found in social networks, news articles and many other document types. Inference of demographic details from social media and online content is useful for marketing, personalization, and forensic purposes and gender prediction has received much attention (Peersman et al., 2011; Argamon et al., 2007).

Shervin Malmasi. 2014. A Data-driven Approach to Studying Given Names and their Gender and Ethnicity Associations. In *Proceedings of Australasian Language Technology Association Workshop*, pages 145−149.

In a study on discriminating gender on Twitter, Burger et al. (2011) used names and screen names as features in their classification system, finding over 400k distinct values for each feature. They found the features to be highly discriminative and informative for this task. Similarly, Tang et al. (2011) take a name-centric approach to gender classification for Facebook, reporting that first names are highly informative for the task.

Name-gender info is also used in the NLP task of co-reference resolution and the state-of-the-art Stanford Deterministic Coreference Resolution System (Lee et al., 2011) uses a list of male and female names to resolve anaphora. However, more generic approaches that use probabilistic models of name features have also been recently applied for this task (Le Nagard and Koehn, 2010).

Name information has also been used in text mining (Patman and Thompson, 2003). One example is in the field of Onomastics where publicly available name information can be used to to infer diversity and gender statistics in various areas.

In computer vision, name information from associated text or captions has been used to aid image-based gender classifiers (Gallagher and Chen, 2008).

3 Data

Due to the paucity of publicly available, machine-readable name data that is annotated for gender we developed and make available our own dataset.

The MQ Names Corpus contains over 13k names from 5 cultural groups, as outlined in Table 1. These include names of Arabic, German, Iranian and Japanese origin. Romanized versions of all names are used. Additionally, a final set of the most common given names sourced from the 1990 US Census data[1] is also included. The data can be obtained by contacting the author.

4 Experimental Methodology

We take a supervised classification approach in this study. More specifically, a linear Support Vector Machine (SVM) classifier is employed.

For features we extract character n-grams of order $n = 1, 2, 3$. These n-grams can help capture orthography, syllables and phonemes. The start and end of names are marked with a special character "$". We do not consider n-grams of higher orders here since these could be capturing whole names instead of more generic patterns.

[1] http://www.census.gov/main/www/cen1990.html

Ethnicity/Culture	Male	Female	Total
Arabic	1090	1148	2238
German	497	576	1073
Iranian	1104	1529	2633
Japanese	1145	1005	2150
US	1219	4275	5494
Total	5055	8533	13588

Table 1: The ethnic groups included in our dataset and the number of names within each gender class.

Baseline (F)	CHAR1	CHAR2	CHAR3
62.8%	67.9%	78.6%	81.3%

Table 2: Gender classification results for the complete dataset using our character n-gram features.

For evaluation, we report our results as classification accuracy under k-fold cross-validation, with $k = 10$. Results are compared with a majority baseline instead of a random baseline as the number of names in each class are unequal.

5 Experiments and Results

In this section we outline our three experiments and present their results.

5.1 Gender Identification

Our first experiment assesses the classification accuracy of gender using our complete dataset.

The results for the baseline (Female) and the 3 feature types are shown in Table 2. All three feature types perform higher than the baseline and character trigrams provide the best performance with 81.3% accuracy. This result demonstrates the presence of gender-predictive features that may generalize across our chosen groups.

5.2 Ethnicity Identification

Our second experiment attempts to predict name ethnicity. 1000 names from each group were used and the ethnic/cultural groups are used as labels. Results are shown in Table 3.

	Baseline	CHAR1	CHAR2	CHAR3
All Cultures	20%	59.8%	72.8%	73.8%
US excluded	25%	70.7%	82.5%	83.5%

Table 3: Ethnicity identification results using 1k names from each group as the data.

While the accuracies across all data are quite high relative to the low baseline, our experiments showed that the US census data performed worse than the other groups. To investigate this we also experimented by excluding the US data. This yields a 10% improvement. In our experiments, all combination of cultures that included the US data performed worse than the combinations that excluded it, so we conclude that this is not just an effect of reducing the number of classes.

We hypothesise that this is because unlike the rest, the US data is sourced from a census, and given the diverse demographics of the US (Hirschman et al., 2000), it is likely to contain names from many ethnicities and cultural groups. For example, there are many names of German origin in the data and this is likely the case

5.3 Gender Classification within Ethnicity

We also experiment with gender prediction within each ethnicity, given that our previous experiment demonstrated that performance may be diminished when data from multiple groups are conflated.

Table 4 includes the results, which show that all languages had higher accuracies than when all the data was combined in experiment 1. Consistent with the results of experiment 2, we also see that the worst performance is on the US data, which is only some 6% above the already high baseline.

6 Feature Analysis

Beyond classification, another interesting task is to analyse the features that distinguish gender and cultural groups. Here, feature ranking could be performed with relevancy methods such as the F-score or Information Gain (Yang and Pedersen, 1997).

However, these methods are limited: they do not provide ranked lists for each gender or ethnicity, but rather an overall rankings. To achieve this, we propose the use of an alternative method using linear SVM weights, as described below.

Using the extracted features, we train linear SVM models for each class (either gender or culture). We use a one-vs-one approach to find features most relevant to each gender. For ethnicity classification, a one-vs-rest approach is used to identify culture-relevant features. L2-regularization is applied to remove noisy features and reduce the candidate feature set size.

Male Names	
Feature	**Examples**
r$	Gunther, Rainer, Heiner
f$	Kristof, Rudolf, Rolf
o$	Ingo, Botho, Waldo
an$	Maximilian, Bastian, Florian
us$	Klaus, Markus, Marius
Female Names	
Feature	**Examples**
a$	Ada, Gisela, Kristina
e$	Heide, Brigitte, Wilhelmine
ild	Hilde, Gerhild, Hildegard
ud$	Gertrud, Hiltrud, Irmtraud
lin	Alina, Karolina, Rosalinde

Table 5: Examples of the highly predictive features for classifying the gender of German names.

Ethnicity	**Features**
Arabic	al ah$ $Mu $Kh $Ab ali yya
German	ert $He ied lin ld$ sch rd$
Iranian	Far eh$ ee$ oo$ ehr okh Gol
Japanese	tsu $Ak ki$ aka suk u$ mi$

Table 6: Examples of features that are highly predictive of each cultural group.

In training the models for each feature, the SVM weight vector[2] is calculated according to (1):

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \quad (1)$$

After training, the positive and negative weights are split into two lists and ranked by weight. The positive weights represent highly predictive features, while features whose absence is indicative of a class will have large negative weights.

We applied this method to our data to extract the most predictive features for classifying the gender of German names. Some of these features, along with sample names containing them, are presented in Table 5. These features highlight that suffixes are an important cue in inferring the gender of German origin names.

We also applied this method to the ethnicity classification problem to extract lists of informative features for each group, shown in Table 6. Looking at the German features we see that they

[2]See Burges (1998) for a detailed explanation.

	Baseline	CHAR1	CHAR2	CHAR3
Arabic	51.3%	71.3% (+20.0)	80.1% (+28.8)	**81.4% (+30.1)**
German	53.7%	73.5% (+19.8)	88.1% (+34.4)	**89.3% (+35.6)**
Iranian	58.1%	68.2% (+10.1)	76.4% (+18.3)	**77.8% (+19.7)**
Japanese	53.3%	72.3% (+19.0)	88.8% (+35.5)	**89.9% (+36.6)**
US	77.8%	79.9% (+2.1)	**84.2% (+6.4)**	82.5% (+4.7)

Table 4: Gender classification results within each ethnic/cultural group of our data. Results include the accuracy and the relative improvement over the baseline. Best results for each group are in bold.

are mostly different than the ones useful for gender prediction.

Another advantage of this method is that it can also produce lists of features that are not associated with each gender or group, although this analysis has been omitted here for reasons of space.

7 Discussion

In this work we demonstrated the utility of a data-driven approach for name research and gender inference, where character trigrams provided the best results. An important finding here is that the gender cues in names are language and culture specific, and classification accuracy is higher when using a model that has been trained with data from the target culture or ethnicity. Since the ethnicity of an unseen name may not be known, this can be predicted first a name-ethnicity classifier, as we demonstrated in experiment 2. Once the most probable cultural group is known, a more specific gender predictor can be applied.

This method can be applied in various areas. It could be used in co-reference resolution systems in cases where unseen names are encountered and their gender must be determined. These methods could also enrich the output of Named-Entity Recognition systems with additional gender information. In text mining and NLP it could be applied to gender and demographic inference tasks for social media and other other big data.

Applications in name research, e.g. Whissell (2001), are also possible. The data-driven feature analysis methodology can help researchers investigating the structure and phonology of names. Additionally, since these gender cues are language-specific, the information extracted here could help non-natives wanting to familiarise themselves with the names of another culture.

The identified features can also be used in other areas. One example is to generate artificial names using features that are hypothesised to be informative about gender, and testing them on subjects to see if they generalize to new cases, e.g. Studies 1 and 2 by Cassidy et al. (1999). They can also be used for the research of product and brand names, which are often not explicitly gendered but strongly associated with customers of a specific gender. In fact, previous findings from applied psychology on phonological cues in product names show that consumers prefer gender congruent names (Cassidy et al., 1999, Studies 7 and 8).

One shortcoming here is the relatively small number of cultural groups included in the data. It would be interesting to assess the task performance with a large number of cultural groups. In the future we plan to expand our dataset with data from more cultural groups including Chinese, Korean, Kurdish, Turkish and Hispanic names. To assist with this, potential sources of publicly available name data need to be identified. Based on our results, data sources where language or ethnicity may not be clearly marked or possibly conflated under another variable, like the US census data, may not be suitable for this task. Wikipedia and public Facebook profiles can be a useful source for name data, although they would need to be annotated. Data sources which include names and gender annotations, such as IMDB, can also help.

Another issue is how to deal with gender neutral names. Unisex names, particularly those varying by culture, can pose problems for our system. One example is the name “Andrea”, which is a feminine name in many cultures but considered a male name in Italy. Additional contextual information, if available, could aid in resolving the appropriate gender.

Another aspect that is how the names are categorized according to “origin”: this could refer to ethnicity, country, language or a cultural group. It may be difficult to categorize all names according to one of these criteria due to confounding factors. One example is the influence of Islam and the Arabic language on Iranian names. Similarly, we also noted that immigration patterns can introduce His-

panic or German origin names into the US census name data. This is an issue that requires further investigation.

Several directions for future research exist. On a basic level, an error analysis could help identify the features causing misclassifications. Another promising avenue is to focus more on the phonetic features of the names by using phonetic encoding techniques such as Soundex or the more advanced Double Metaphone algorithm (Philips, 2000) to transform the representations of the names.

Another direction is to expand this research to surnames, which are also known to have their own unique structures and meanings (Reaney, 1991). Here, data-driven machine learning techniques could be applied to large-scale data to examine how patronymic, toponymic and locative aspects of last names from different cultures could be learnt algorithmically.

In sum, the results from this research, an intersection of linguistics, psychology and machine learning, can help inform name research and be applied in a variety of areas.

Acknowledgments

We would like to thank the reviewers for their insightful feedback and constructive comments.

References

Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2007. Mining the Blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9).

Herbert Barry Jr and Aylene S Harper. 1995. Increased choice of female phonetic attributes in first names. *Sex Roles*, 32(11-12):809–819.

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics.

Christopher JC Burges. 1998. A tutorial on Support Vector Machines for Pattern Recognition. *Data mining and knowledge discovery*, 2(2):121–167.

Kimberly Wright Cassidy, Michael H Kelly, and Lee'at J Sharoni. 1999. Inferring gender from name phonology. *Journal of Experimental Psychology: General*, 128(3):362.

Anne Cutler, James McQueen, and Ken Robinson. 1990. Elizabeth and John: Sound patterns of men's and women's names. *Journal of linguistics*, 26(02):471–482.

Andrew C Gallagher and Tsuhan Chen. 2008. Estimating age, gender, and identity using first name priors. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

Charles Hirschman, Richard Alba, and Reynolds Farley. 2000. The meaning and measurement of race in the US Census: Glimpses into the future. *Demography*, 37(3):381–393.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics*, pages 252–261. Association for Computational Linguistics.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34. Association for Computational Linguistics.

Frankie Patman and Paul Thompson. 2003. Names: A new frontier in text mining. In *Intelligence and Security Informatics*, pages 27–38. Springer.

Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

Lawrence Philips. 2000. The double metaphone search algorithm. *C/C++ users journal*, 18(6):38–43.

Percy Hide Reaney. 1991. *A dictionary of English surnames*. Psychology Press.

Anne Saxon Slater and Saul Feinman. 1985. Gender and the phonology of North American first names. *Sex Roles*, 13(7-8):429–440.

Cong Tang, Keith Ross, Nitesh Saxena, and Ruichuan Chen. 2011. What's in a name: A study of names, gender inference, and gender behavior in Facebook. In *Database Systems for Adanced Applications*, pages 344–356. Springer.

Cynthia Whissell. 2001. Sound and emotion in given names. *Names*, 49(2):97–120.

Yiming Yang and Jan O Pedersen. 1997. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420.

Parvin Yavari, T Gregory Hislop, Chris Bajdik, Alireza Sadjadi, Mehdi Nouraie, Masoud Babai, and Reza Malekzadeh. 2006. Comparison of cancer incidence in Iran and Iranian immigrants to British Columbia, Canada. *Asian Pacific Journal of Cancer Prevention*, 7(1):86.