Names | A Journal of Onomastics

Corpus-Based Methods for Recognizing the Gender of Anthroponyms

Rogelio Nazar, Irene Renau, Nicolás Acosta, Hernán Robledo, Maha Soliman, and Sofía Zamora

Instituto de Literatura y Ciencias del Lenguaje Pontificia Universidad Católica de Valparaíso, Valparaíso, CHILE

This paper presents a series of methods for automatically determining the gender of proper names, based on their co-occurrence with words and grammatical features in a large corpus. Although the results obtained were for Spanish given names, the method presented here can be easily replicated and used for names in other languages. Most methods reported in the literature use pre-existing lists of first names that require costly manual processing and tend to become quickly outdated. Instead, we propose using corpora. Doing so offers the possibility of obtaining real and up-to-date name-gender links. To test the effectiveness of our method, we explored various machine-learning methods as well as another method based on simple frequency of co-occurrence. The latter produced the best results: 93% precision and 88% recall on a database of ca. 10,000 mixed names. Our method can be applied to a variety of natural language processing tasks such as information extraction, machine translation, anaphora resolution or large-scale delivery or email correspondence, among others.

Keywords: anthroponymy, co-occurrence statistics, corpus linguistics, gender recognition, given names, Spanish

ans-names.pitt.edu

ISSN: 0027-7738 (print) 1756-2279 (web) Vol. 69, Issue 3, Summer 2021 DOI 10.5195/names.2021.2238

1. Introduction

In this paper, we address the use of corpora for the automatic recognition of the gender of proper names, specifically anthroponyms. Automatic detection of the gender of names is very useful for a variety of tasks. It allows researchers to obtain a large amount of empirical data without depending on manually compiled lists. It can also be used to solve problems related to anaphora resolution, a challenge in different natural language-processing tasks such as machine translation. For example, knowing the gender of the name is a crucial step in solving grammatical agreement correctly, especially when the target language has gender features marked by pronouns or adjectives. More broadly, it can be applied to large-scale studies about gender bias or other social issues. As this investigation will demonstrate, this work can also have practical applications.

In the present study, the focus is name gender detection for the large-scale delivery of emails which is commonly needed by private and public organizations. For example, university professors are engaged in frequent email correspondence with students and could save valuable time by automating at least part of the process. In such automation, people must be addressed by name, either by the first name, the last name, or both. And this is where the automatic detection of gender has a role to play, because the generated messages have to observe the morphological markings of gender that are present in most languages. In all Romance languages, as well as languages from other families such as German, Urdu, Korean, and Dyirbal, feminine and masculine endings of nouns and adjectives, as well as masculine and feminine pronouns, are very frequent. And even in languages where gender is not morphologically marked, such as English, recipients are nonetheless addressed by their titles (*Mr.*, *Mrs.*, etc.).

In commercial correspondence, not knowing the gender of the addressee has led on many occasions to sub-optimal solutions. One of them has been using only the masculine form, as in *Estimado Irene 'Dear Irene,' where a masculine adjective estimado "dear" is used to refer to a feminine name. Another case that is commonly observed is to offer both the masculine and feminine endings, e.g., *Estimado(a) cliente(a)*. From a prescriptive point of view, the first option is grammatically incorrect, and the second one may be considered by some to be pragmatically inadequate because it makes the message look impersonal. Both alternatives can damage the personalization of the message (specifically, the identification strategy), which is a key aspect of the communication between a company and their customers (Maslowska, Smit, and van den Putte 2016). More advanced solutions to this problem have been reported in the literature, such as automatic name-gender recognition (Naldi et al. 2005; Lax Martínez, Raffo, and Saito 2016). However, most of the proposals use already existing lists of first names. Some of the drawbacks of such an approach are the difficulty of manually compiling such lists, the problem of resources becoming outdated, and the challenge of analyzing novel names.

In this article, we report on the results of different methods that predict the gender of a target name by analyzing the co-occurring words in large corpora. We tested different machine learning methods as well as one based on simple frequency of co-occurrence. Contrary to our expectations, it was the latter one that produced the best results: 93% precision and 88% recall on a database of 10,000 mixed names. We used a Spanish corpus for the study. However, explicit knowledge of this language is minimal as the material used consists only of short word lists.

In this study, we define the gender of anthroponyms linguistically. In languages with grammatical gender, all common nouns and proper names have gender. For example, *mesa* "table" in Spanish is feminine; and in German, *Tisch*, is masculine. In the case of anthroponyms, the formal properties of the grammatical gender are inherited such that complementary parts of speech correspondingly adopt grammatical gender (Motschenbacher 2020). Hence, in our proposal, we do not consider biological, psychological, or social gender, which cannot be detected with a purely corpus linguistics approach. Accordingly, our method is not able to detect whether the name *María* refers to a man or a woman. The method can only determine that *María* is a feminine anthroponym.

2. State of the Art in Predicting the Gender of Anthroponyms

Although the automatic detection of the gender of anthroponyms has been studied by groups of researchers focusing on various languages and with different purposes, previous work on the subject is not abundant. The most common strategy consists of obtaining gender information of names from commercial or government sources. This is the method that is most commonly applied by authors interested in the detection of gender bias in scientific and technological publications (Naldi and Parenti 2002; Naldi et al. 2005; Frietsch et al. 2009; Kugele 2010; National Women's Business Council 2012; Larivière et al. 2013; Sugimoto et al. 2015; Lax Martínez, Raffo, and Saito 2016).

Naldi and Parenti (2002) and Naldi et al. (2005) used different sources, such as "dictionaries, calendars, books and internet sites, files from Record Offices, and phone books" (Naldi et al. 2005, 304). They compiled a First Name Database containing 8,291 different names with associated information about gender, covering six European languages (English, French, German, Italian, Spanish, and Swedish). Using a similar strategy, Larivière et al. (2013) compiled an extensive list of names associated with gender by collecting already existing lists such as different national censuses, Wikipedia universal lists of names, and others. Lax Martínez, Raffo, and Saito (2016) used 13 sources provided by national institutions covering 173 countries and collected around 394,422 unique names classified by gender. There are also some smaller scale studies that use only one source to study tendencies regarding names, such as in Barry and Harper (2014) on unisex names. In addition, Tang et al. (2011) used Facebook's user information to gather a list of names and their genders, and combined them with methods regarding author gender prediction.

The approaches based on data collection have the limitation that, no matter how extensive these lists are, they are static and not sensitive to the creation of new names and the arrival of those due to migratory movements, as shown by Gao (2011), Parada (2016), Giménez (2017), and Sabet and Zhang (2020), who described the cultural, social, and ideological circumstances affecting the dynamics of name change. Thus, the lists of names would have to be constantly updated in order to remain useful, and this would of course be very costly. This is why an alternative approach has to be taken into account, and that is to pose the problem of name-gender prediction as a classification task, independently of any data collection. This way, names can be analyzed regardless of their inclusion in manually collected data lists.

Regarding the design of algorithms to predict the gender of names, some software solutions have already appeared on the market, such as Genderize.io (https://genderize.io/) or Gender-Api (https://gender-api.com/). However, no details about the performance and design of these systems are publicly available, and there are no elements to know if they are doing anything other than looking up an already compiled database of names. In the NLP community, in turn, research in this particular area is rather scarce. Among the few papers available, we find the study by Tripathi and Faruqui (2011), who predict the gender of names in languages spoken in India (Hindi, Urdu, Tamil, among others) and used a combination of features such as character length, number of syllables, and vowel ending among others. They experimented with a list of 2,000 names and used a support vector machine classifier (SVM) for the classification. Ali et al. (2016) applied a similar strategy for Urdu names written with the Latin alphabet, relying on features such as name length and character sequences.

Park and Yoon (2007) dealt with the task of gender identification with the purpose of improving Korean-English machine translation. In Korean, subject or object pronouns can be omitted and this affects anaphora resolution. This means that one needs to know the gender of the omitted subject/object in order to produce a correct translation. For the classification, they applied a machine-learning algorithm that uses sequences of three syllables as features. An approach more in line with our present study was adopted by Yoon et al. (2008) who also used machine learning but extended the research by Park and Yoon (2007) to include not only features of the name they are trying to classify but also of words that co-occur with said names. As they observe, some words tend to co-occur with names of one gender rather than the other. In our present research, we further develop the idea of using contextual features to predict the gender of a name, but we explore the use of simpler methods which can be faster, less computationally expensive, and more easily adapted for languages other than Korean.

3. Methodology

The methodological approach we used to predict the gender of names is based on word co-occurrence in a large textual corpus. Our method consists of measuring the frequency of words co-occurring with the target name in a symmetrical ten-word context window. This window size represents the best balance because a shorter one would be appropriate only for the study of different types of phenomena, such as collocations of multi-word terminology, which consist of combinations of adjacent or semi-adjacent elements. In contrast, the type of co-occurring units we are interested in can appear at larger distances. However, a window larger than that would incorporate more material unrelated to the target name and could potentially skew results.

We therefore defined a set $X = X_{1,} ... X_n$ as input, consisting of a list of masculine and feminine proper names. The output would then be a gender value for each name x. Thus, for all x in set X there is a "gender" function G: $(\forall x \in X)G(x)$. The function returns one of three possible values: $\{M, F, U\}$ for *masculine*, *feminine*, and *undefined*, the latter being reserved for the cases in which not enough information is available to assert a gender with confidence.

As far as we know, this is a novel approach to the problem of automatic detection of the gender of names. In comparison with existing strategies, the one we propose has the advantage of the flexibility provided by language use, which is reflected in the corpus, and allows for easy and fast updating of the data. The closest approach is the already-mentioned work by Yoon et al. (2008), who predict the gender of a name based on the words it co-occurs with. But theirs is a language-specific and very complex approach, as they select the co-occurring words by an information-entropy model and then use an SVM machine-learning approach that requires a Korean morphological analyzer. In contrast, ours is a minimalistic approach which requires no sophisticated feature selection.

For our experiments, we use a Spanish corpus to detect the gender of names of any language, due to the fact that names are relatively stable across languages. For instance, a text written in Spanish about William

Faulkner would most likely not refer to him as *Guillermo Faulkner*. For this simple reason, we can exploit the gender markedness of a language like Spanish to predict the gender of English names. Of course, we could further extend our method by including corpora of different languages with grammatical gender, such as French, German, etc., but we restricted the materials to Spanish in order to keep the procedure as simple as possible. In any case, and despite the fact that ours is not strictly speaking a language-independent approach, the method is simple and fairly easy to replicate in a different language.

The corpus we used is the EsTenTen (Kilgarriff and Renau 2013), a large sample of text in Spanish (9,526,603,050 tokens, not counting punctuation marks) downloaded from the web. It is organized in sections corresponding to different Spanish-speaking countries, but for our experiments we limited the sample to Argentina, Chile, Colombia, Mexico, and Spain.

The method we propose can be applied in isolation or combined with existing lists, for example, to update a resource. The natural setting for the application of our method would be to add a gender field to a database of names. A normal format in a database containing names of people is to have the first name plus a middle name, if any, followed by the last name, and this is how we formatted our input data for our experiments. We must clarify, however, that the classification is conducted using only the given names of the input list. Therefore, we assume hereafter that the symbol x denotes only the first name, not the family name, which is irrelevant for gender prediction. Once we have attributed a gender to each given name in the input list, classifying each full name is then straightforward: if *María* and *Isabel* are feminine names, then *María Isabel Gómez* is a woman's name. There are only few exceptions, like *María José* and *José María*, the first being a feminine name and the second a masculine name in Spanish. However, cases like these are statistically insignificant and could be dealt with using a specific rule to capture the order in which they appear.

3.1. The Basic Algorithm

The basic algorithm is the simplest method we devised for the classification of given names by analyzing the contexts of occurrence of name x in a large corpus. Our hypothesis was that elements co-occurring in the context of occurrence of x will work as predictors of x's gender. As already mentioned, we use a symmetric context window of ten words. Elements working as estimators for the gender prediction are defined as lexical units (nouns or adjectives) and grammatical features (articles) found at the left or right of the target name x within the context of occurrence. Consider, for illustration, the examples of the name *Cindy*, in Table 1. Elements marked in bold are those which can be used as gender predictors of the name.

Table 1. Examples of Concordances of the Name *Cindy* in the EsTenTen Corpus

el repo, el reportero hizo buenas migas con	Cindy	**, una** simpática y dócil mona que parecía
de la Universidad de Austin, y	Cindy	Meston**, directora** del Laboratorio de
de ciertas empresas de cosméticos,	Cindy	Crawford fue demasiado **vieja** a los 35
que nos quedó pendiente Victor hermosa	Cindy	**una niña** hermosa que quedó
no decir" Esta noticia surge mientras	Cindy	McCain**, la esposa** del senador
, una se llama Gina y **la otra**	Cindy	Love, muy parecidas y casi confundibles

As illustrated in the table above, one linguistic feature that is indicative of gender are the articles, particularly when used after a comma. In this case, the personal name *Cindy* is accompanied by the feminine form (*una* 'a'). Other indicative elements are adjectives in grammatical agreement. In the example above, there are many adjectives with feminine inflection, such as *simpática* 'nice', *vieja* 'old', and *hermosa* 'beautiful'. There are also co-occurring gender-marked nouns such as *niña* 'girl', *esposa* 'wife', etc.

The general idea of the basic algorithm is to compile lists of words associated with each gender. Searches for these words are then performed within the contexts of occurrence of name x and used as gender predictors. We derived these lexical units first from an arbitrary selection and later by inductive methods using corpora as explained in section 3.2. Irrespective of the origin of these features, we divided them into two sets, M and F (*where* $|M| = |F|$), and for any input name x, we proceeded to extract a set C of randomly selected corpus concordances ($|C| \leq 5000$). It is important to stress that we predict the gender of a name after analyzing the thousands of contexts of such name in the corpus. If x = *Juan*, it is not the case that we are predicting the gender of this name after a particular occurrence of said name in a single context. This is why it does not matter if in one or more contexts we find that the name *Juan* appears with other names of the opposite gender, as in "Cindy y Juan" or "Juan y María." The set of contexts to analyze in such case will still be those of *Juan*, and the occurrence of other names will not affect the overall counts for the features co-occurring with *Juan*.

The program we implemented loops through the contexts of occurrence of x and checks for the occurrence of any element in feature sets M and F. Formally, we defined a context of occurrence of x as a lexical/function word set $C_j = \{t_1, \ldots t_{|Cj|}\}$, therefore ignoring word order. Each lexical or grammatical unit

$t \in Cj$ is a word occurring at either side of target name x. We defined an estimator, $K_i(x)$, with two possible values (i ∊ 2 {f , m}). There is a masculine estimator, K_m, for x in (I), which yields an integer value representing the number of times some feature of set M was found within all the contexts of x. Conversely, an identical estimator $K_f(x)$ is defined for the case of the feminine gender using set F instead of M.

$$K_m(x) = \sum_{j=1}^{|C|} \begin{cases} 1 \; if \; (\exists t \; \in \; C_j^t \; \in \; M) \\ 0 \quad \text{otherwise} \end{cases}$$

After all contexts of occurrence have been analyzed, a decision is made with function $G(x)$ to classify input name x either as M, F, or U, according to Equation 2. The value U, for undefined, means that the test is inconclusive because there were not enough contexts of occurrence of x in the corpus, as expressed in the first condition, where u is an arbitrary threshold (set to 10 on an empirical basis).

$$\mathrm{G}(x) = \begin{cases} \mathrm{U} \quad \text{if} \, |\mathrm{C}| \leq \mathrm{u} \\ \mathrm{M} \quad \text{if} \, K_m(x) > K_f(x) \\ \mathrm{F} \quad \text{otherwise} \end{cases}$$

The rationale behind this strategy is that masculine names tend to co-occur more often with words that are associated with men rather than with women. A different problem concerned determining which features should be used as gender estimators, as we explain in the following section.

3.2. Feature Selection

As is customary in classification algorithms, we use the term “feature” to denote any trait that can be used for the classification. That means it not only denotes “morphological features” because, in this sense, lexical units can be features as well. As already mentioned, the basic algorithm to predict the gender of proper names used an arbitrarily selected list of features. These can be not only morphological features proper, but also words, sequences of words, and sequences of punctuation signs with words. These elements are expected to co-occur in the linguistic contexts of a name and in agreement with its gender. We experimented with two versions of such a list. The first one was very limited in number, with only two features as seen below in Table 2. The second list of features was slightly extended, as shown in Table 3.

Table 2. Short List of Arbitrarily Selected Features

Gender	Original list of features in Spanish
Feminine	es una ‘is a [fem]’ /, una ‘, a [fem]’
Masculine	es un ‘is a [masc]’ /, un ‘, a [masc]’

Table 3. Extended List of Features Selected by Introspection

Gender	Original list of features in Spanish
Feminine	es una ‘is a[fem]’ /, una ‘, a[fem]’ / una joven ‘a young woman’ / chica ‘girl’ / señora ‘lady’ / señorita ‘miss’ / actriz ‘actress’ / esposa ‘wife’ / mujer ‘woman’ / madre ‘mother’ / hermana ‘sister’ / hija ‘daughter’ / doña ‘miss’ / sra. ‘mrs.’
Masculine	es un ‘is a[masc]’ /, un ‘, a[masc]’ / un joven ‘a young man’ / chico ‘boy’ / señor ‘sir’ / actor ‘actor’ / esposo ‘husband’ / marido ‘husband’ / padre ‘father’ / hermano ‘brother’ / hijo ‘son’ / don ‘mr.’ / sr. ‘mr.’

The features were chosen upon simple introspection. The guiding criterion for making the selection was identifying words that would typically co-occur with names of each gender. In the case of the short list, the features most reliably associated with the gender were the gender-marked indefinite articles (*un* ‘a’ for the masculine gender and *una* ‘a’ for the feminine) after a comma or semicolon. This feature set made it possible to retrieve text passages like the first pattern in Table 1: *Cindy, una simpática y dócil mona* “Cindy, a nice and docile monkey”.

The extended list contains only a few more nouns (27 in total) expected to be indicators of gender. In particular, nouns that indicated the profession, kinship role, or marital status of the name-bearer were selected. In this regard, we selected feminine lexical features such as *mujer* 'woman', *señora* 'lady', and *señorita* 'Ms.'. Similarly, the same procedure was followed for the masculine lexical features. This was the case of *señor* 'Mr.', *marido* 'husband', and *don* 'Sir', as well as others such as *actor* 'actor' and *actriz* 'actress', as shown in Table 3.

In addition to the list of features we created by introspection, we also attempted to automatically extract features from the corpus using an inductive process. In order to do that, we drew a random sample of 100 masculine and 100 feminine names (i.e., names of which we already knew the gender) and then retrieved 5,000 random contexts of occurrence per name, thereby obtaining a sample of 500,000 contexts of occurrence of men's names and another sample of the same size of women's names. Again, as before, the size of the context window is ten words to the left and right. From each of these sub-corpora, we extracted a list of all single words appearing in these contexts, sorted by decreasing order of frequency. We then divided these lists of words occurring in the contexts into a short and an extended version. The short version was obtained by including the 500 most frequently occurring words. We removed the intersection between both lists in order to discard those units that would not be helpful for the categorization precisely because they co-occur frequently with both genders. These are non-informative, high-frequency words (function words) such as *también* 'also', *cuando* 'when', or *desde* 'since'. As a result, we obtained two lists with 87 words for each gender. For the extended version, we followed the same procedure, but this time retaining the 1,500 most frequently occurring words. This resulted in 303 words for each gender.

In this inductive phase, features that were selected according to their frequency of co-occurrence in corpus also provide information about the marital status of the person, as was the case for those we selected by introspection. Accordingly, inductive features include details about the subjects' nationality, such as *alemán* 'German' [masc.] or *británico* 'British' [masc.], as well as their professions, e.g., *ministro* 'minister' [masc.] or *directora* 'director' [fem.]. These elements carry morphological information about the gender as some co-occurring adjectives such as *famoso* 'famous' [masc.] and *nueva* 'new' [fem.].

3.3. Machine-Learning Methods

In order to contrast the performance of the basic algorithm with other approaches, we opted for the use of machine-learning algorithms. We find this categorization problem suitable for such methods, because we can train a classifier to use examples of masculine and feminine names, using the same type of features as those described in the previous subsection. We experimented with some of the most well-known supervised automatic classification algorithms, such as Naive Bayes (Maron 1961; Mosteller and Wallace 1964), the J48 decision tree algorithm (Quinlan 1993), Support Vector Machine (SVM, Vapnik 1998), and Sequential Minimal Optimization (SMO, Platt 1998). For convenience, we used the implementation offered by Weka (Hall et al. 2009), the popular software platform for machine learning. As features for these classifiers, we used the same lists as with the first algorithm, in order to have a meaningful comparison. These features are represented in a matrix for each instance of our collection. In order to diversify possibilities, we developed two kinds of matrices: one with binary values, and one with real-value numbers. In the first case, the matrix represents only the presence or absence of a given feature *i* on name *j*. In the other case, the values represent the frequency of co-occurrence of feature *i* and name *j*.

With these classifiers and matrices in hand, we proceeded to perform eight rounds of experiments for each of the four methods (i.e., the four short/long deductive/inductive combinations). We then compared the results with the four rounds of experiments we had conducted using the basic algorithm.

The experiments with machine-learning methods were conducted using the ten-fold validation method. This meant each dataset was divided into ten parts and the experiment was conducted ten times—each time using 90% of the dataset for training and 10% for testing; then averaging the figures of precision and recall for all ten experiments.

4. Results

For our experiments we used lists of names taken from Wikipedia, as they are classified by gender.[1] Another advantage of these lists is that they represent a large enough sample to obtain reliable estimations of performance of a classification algorithm. It should be stressed that we only used these lists to test the method, not as part of the method. From these lists, we obtained the names of 4,790 women (e.g., *Edith Lagos*, *Federica País*, *Linda Indergand*, etc.), and 5,075 men (e.g., *Pascual Baburizza*, *Jonathan Fabbro*, *Imil Habibi*, etc.). With this list of 9,865 names, we performed the experiment, first using the basic algorithm and then using each of the machine-learning methods in their different parameter combinations.

4.1. Comparative Figures of Precision, Recall, and F1

In order to assess the performance of our method, we used the evaluation measures of precision, recall, and F1, which are commonly used in Natural Language Processing (Manning and Schütze 1999). "Precision" refers to the proportion of correct results among the output of the algorithm. "Recall," in turn, refers to how exhaustive the method is. Finally, the "F1" score represents a balance between the above two values.

Tables 4–7 present the values of precision, recall, and F1 for each of the 36 experiments we conducted. According to these figures, the best performing classifier in terms of precision was the basic algorithm, yielding 93% with the long deductive setting (Table 5). Likewise, this algorithm produced the best recall, returning 92% with the short deductive setting (Table 4). In terms of the F1-score, however, the basic algorithm, at its best performing setting, was superseded by the J48 Decision Tree algorithm working with the real value frequency matrix. However, there was not a significant margin—90% against 91% (Table 5). Contrary to our expectations, the experimental sets using the inductive features did not outperform those using the deductive ones (Tables 6 and 7); and the best performing algorithms, SMO and J48, did not surpass the limit of 90% in terms of precision, recall, and F1. Interestingly, in the case of the long inductive setting (Table 7), SMO performed better with the binary matrix than it did with the real value matrix.

Table 4. Experimental Results Using the Short Deductive Setting

Experiment	Precision	Recall	F1
Basic algorithm	0.83	0.92	0.87
Binary Naive Bayes	0.49	0.50	0.50
Frequency Naive Bayes	0.73	0.70	0.70
Binary SVM	0.49	0.57	0.52
Frequency SVM	0.88	0.87	0.87
Binary SMO	0.50	0.53	0.51
Frequency SMO	0.84	0.81	0.80
Binary J48	0.49	0.57	0.52
Frequency J48	0.90	0.90	0.90

Table 5. Experimental Results Using the Long Deductive Setting

Experiment	Precision	Recall	F1
Basic algorithm	0.93	0.88	0.90
Binary Naive Bayes	0.66	0.62	0.60
Frequency Naive Bayes	0.79	0.70	0.68
Binary SVM	0.70	0.62	0.58
Frequency SVM	0.88	0.87	0.87
Binary SMO	0.69	0.62	0.58
Frequency SMO	0.87	0.84	0.84
Binary J48	0.67	0.63	0.61
Frequency J48	0.91	0.91	0.91

Table 6. Experimental Results Using the Short Inductive Setting

Experiment	Precision	Recall	F1
Basic algorithm	0.69	0.79	0.73
Binary Naive Bayes	0.56	0.54	0.51
Frequency Naive Bayes	0.81	0.77	0.76
Binary SVM	0.72	0.66	0.64
Frequency SVM	0.87	0.85	0.85
Binary SMO	0.73	0.68	0.66
Frequency SMO	0.87	0.85	0.84
Binary J48	0.71	0.67	0.66
Frequency J48	0.90	0.90	0.90

Table 7. Experimental Results Using the Long Inductive Setting

Experiment	Precision	Recall	F1
Basic algorithm	0.66	0.76	0.70
Binary Naive Bayes	0.57	0.56	0.53
Frequency Naive Bayes	0.82	0.78	0.78
Binary SVM	0.89	0.89	0.89
Frequency SVM	0.87	0.85	0.85
Binary SMO	0.90	0.90	0.90
Frequency SMO	0.88	0.86	0.86
Binary J48	0.88	0.88	0.88
Frequency J48	0.90	0.90	0.90

4.2. Error Analysis

A qualitative analysis of errors (541 in total) offers valuable insight into their causes and suggests possible ways to reduce them in the future. We conducted an error analysis on the results of the best setting of the basic algorithm, i.e., the long deductive method that consisted of the longer list of arbitrarily selected features. The following two tables show samples of cases of wrongly classified names. Table 8 presents some of the men's names that were mistakenly classified as women's names, while Table 9 shows the opposite. In both tables, cases marked with "*" are instances where the algorithm was actually correct, and the error was in the gold-standard (0.9% of the total).

Table 8. Examples of Names of Men Incorrectly Classified as Women by the Basic Algorithm with the Best Performing Setting (Long Deductive Method)

Acie Earl, Alaeddine Yahia, Amata Kabua, Ana María De Saavedra y de Macia,, Andrea Bacchetti, Bertus Aafjes, Blago Zadro, Bõstjan Nachbar, Braj Kachru, Bülent Ecevit, ***Carla de Sa**, ***Carmen Tagle**, Celestine Babayaro, Chico Xavier, Christy Cabanne, Cleó, Danai Udomchoke, Daryl Sabara, Demonaz Doom Occulta, Desean Jackson, Diosdado Macapagal, Dominique Da Sylva, Dragomir Racic, Ehsan Hadadi, Estácio De Sá Eunan OʼKane, Faissal Ebnoutalib, Flann OʼBrien, Fons Rademakers, Gehad El-Haddad, Ghasem Hadadifar, ***Giovanna Valcárcel**, Hanibal Al Gadafi, Harley Earl, Haruna Babangida, Heiki Nabi, Hotaru Yamaguchi

Table 9. A Sample of Names of Women Incorrectly Classified as Men by the Basic Algorithm with the Best Performing Setting (Long Deductive Method)

Abella Danger, Africa Zamorano, Ainsley Bailey, Ala Baguiyants, Ali Macgraw, Alona Tal, Aly Wagner, América Valenzuela, Ander Page, Annalena Mcafee, Aria Wallace, Ariel Kaplan, Armi Aavikko, Arrate Egaña, Arzu Tan, Astra Zarina, Ataru Nakamura, Atena Farghadani, Attiya Inayatullah, Avelina Valladares, Azar Nafisi, Babe Paley, Badia Hadj Nasser, Bak Ji-Yun, Bako Dagnon, Bella Paige, Beren Saat, Binnie Hale, Blessing Oborududu, Brandy Talore, Bristol Palin, Brody Dalle, Cai Wenji, ***Carlos Eduardo Zavaleta Rivera**, Casey Labow, Cele Abba, Chaitra H. G., Chitose Hajime, Christian Bach, Ciaran Madden, Claude Jade, Coral Herrera, Cox Habbema, Dad Dáger, Daini No Sanmi, Dai Qing, Daisuke Higuchi, Dale Raoul, Dami Im, Dee Edwards, Deja Daire

An examination of the misclassifications revealed that a frequent source of error (9.4%) were names which are usually feminine in one language but masculine in another (e.g., *Andrea, Charlie*, *Simone*). Luckily, there are a comparatively limited number of names that account for this kind of error. One solution to this problem might be the use of machine learning that would associate family names with nationalities, in order to derive rules such as $GivenName = Andrea \wedge Surname \in \{ItalianNames\} \rightarrow G(Andrea) = M$. This does not mean that countries are necessarily monolingual and/or monocultural, as they often present mixed cultural heritage and names. Nevertheless, we contend that such a methodological modification would help reduce the error rate by some margin, because, to continue with the same example, most people called *Andrea* in Italy will be men.

Another source of errors involved unisex names (2.2%). We believe that solving this problem would require methods different than the one proposed in this article. It is, indeed, an attractive topic for future research, but we can already imagine how it could be done: to determine the gender of *Ariel* or *Jean*, we would first need to confirm whether it is a unisex name. If this were the case, the next step would be to search for contexts of occurrence of the full name of the person, e.g., *Ariel Smith* or *Jean Smith*. This means that the approach would be focused on attempting to predict the gender of that specific person and not of the name in general, as we have done in this paper. However, such a method would be useful only if there were sufficient text written about such individual, as is the case with public figures.

To a lesser extent (0.3%), we encountered errors with names that have the same form as other words with different grammatical gender. This was the case of *Sol*, which is not only a woman's name in Spanish, but is also the word for sun, a noun with masculine gender. This error, however, only happened when names coincided with common nouns, such as the female names *Dolores*, *Pilar*, *Rosario*, and the masculine common nouns *dolores* 'pains,' *pilar* 'post,' and *rosario* 'rosary.' Another infrequent (0.5%) type of error concerned instances of masculine names adopted as pen names by women writers, such as *Fernán Caballero* or *George Egerton*. We must, however, argue here that although this counts as an error in the evaluation, technically speaking it is not a real error of the method. As we already stated, the purpose of the experiment is to predict the gender of the name, not the gender of the person.

From all the different types of errors we found, the most frequent (73.4%) involved names that are too infrequent in the analyzed language, Spanish. The vast majority of the examples of errors shown in Tables 8 and 9 are cases of non-Spanish names that are very infrequent in the corpus. When names are very infrequent in the corpus, that means that the algorithm has few contexts of occurrence from which to compute the statistics, and therefore results are less reliable because they are more sensitive to random occurrence of elements. Probably the best way to mitigate this problem would be triangulation using multilingual corpora, specifically with languages that observe grammatical gender. For example, having also a French and a German corpus would increase the chances of finding contexts of occurrence of such uncommon names. Finally, 13.3% of the errors were due to a miscellany of other reasons.

5. Conclusions

In this paper, we examined the problem of predicting the grammatical gender of names and proposed an algorithm to solve said problem on the basis of corpus statistics. We did this by analyzing words that frequently co-occur with a given name and are deemed reliable predictors of name gender. We explored this potential solution using different sets of predictors: first with only determiners; then with an extended list of arbitrarily selected features; and lastly with features inductively obtained from the corpus. In order to assess the effectiveness of the proposed algorithm, we also compared the performance of the classifier with a series of well-known machine-learning algorithms, using the same features and dataset.

It was surprising to discover that the basic algorithm, as simple as it is, was able to perform on a level comparable to that of the more sophisticated machine-learning algorithms. Of course, when different algorithms produce similar results, there is a natural preference for the most parsimonious one, because often the more complex an algorithm is, the more execution time is required. In addition to saving time, the basic algorithm also performed the best in terms of precision (93%) and recall (92%), although not in F1. Given the fact that the envisioned application of this algorithm is the large-scale distribution of email correspondence, we favor more conservative approaches such as the basic algorithm to reduce errors of gender assignment. Accomplishing this task quickly and reliably is more important than achieving exhaustiveness.

We have implemented our algorithm on a web platform (http://www.tecling.com/genom) that receives a list of names and returns the same list with a new column indicating the gender (*M, F,* or *U*). Although the platform accepts names in any language, it performs best with names in Spanish since at the moment, it only retrieves contexts from a Spanish corpus. This, however, does not hinder the replication of our method using other languages, including English, despite the fact that this language lacks gender morphology. There are two reasons why this is not a hindrance. The first is that in English, there are also lexical units associated with gender (e.g., *father, mother, brother, sister*). The second is that one can use a French or a Spanish corpus to process English names, due to the stability of names across languages commented upon in Section 3.

There are many potential applications of the proposed method in various tasks of natural language processing. Examples are information extraction and machine translation, because knowing the gender of names can play an important role in anaphora resolution, which is a key part of those tasks. We also foresee how the method could be applied in large-scale studies that investigate gender bias. The most immediate and concrete application, however, is the one we mentioned at the beginning of this paper—to improve the large-scale delivery of email correspondence by correctly addressing the recipients by gender.

Many directions for future research may evolve from this investigation. We already mentioned some ideas for the case of unisex names. It would be interesting to further investigate how unisex names are related to language, culture, and nationality. For instance, *Jean* is quite often considered a feminine name in English but a masculine one in French. In the same way, it would also be interesting to determine whether a similar relationship holds between different dialects and varieties of the same language. Finally, yet another intriguing idea would be to study how unisex names are correlated with different generations. Clearly, there are many fascinating questions which have yet to be explored in this area of research.

Note

1. The lists of masculine and feminine names for our experiments were obtained from the following Wikipedia pages:

https://es.wikipedia.org/wiki/Categor%C3% ADa:Hombres

https://es.wikipedia.org/wiki/Categor% C3%ADa:Mujeres

Acknowledgments

The authors would also like to thank the reviewers and the editor for the many helpful comments and insights.

Funding

This research was made possible thanks to two successive grants: project 035.691-32, funded by Pontificia Universidad Católica de Valparaíso (2017), and project Fondecyt Regular 1191481 (2019–2021), funded by ANID (Chile Government).

Bibliography

Ali, Daler, Malik Muhammad Saad Missen, Nadeem Akhtar, Nadeem Salamat, Hina Asmat, and Amnah Firdous. 2016. "Gender Prediction for Expert Finding Task." *International Journal of Advanced Computer Science and Applications* 7, no. 5: 161–5.

Barry, Herbert, III, and Aylene S. Harper. 2014. "Unisex Names for Babies Born in Pennsylvania. 1990–2010." *Names* 62, no. 1: 13–22.

Frietsch, Rainer, Inna Haller, Melanie Vrohlings, and Hariolf Grupp. 2009. "Gender-Specific Patterns in Patenting and Publishing." R*esearch Policy* 38, no. 4: 590–9.

Gao, Ge. 2011. "Shall I Name Her 'Wisdom' or 'Elegance'? Naming in China." *Names* 59, no. 3: 164–74.

Giménez, Iván. 2017. "Nombres de bebés, bares, viajes ... la locura desatada por *Juego de Tronos*." [Names of babies, bars, trips ... the madness unleashed by Game of Thrones]. *La Vanguardia*, June 20, 2017. Accessed May 5, 2020. http://www.lavanguardia.com/series/20170720/424201948012/juego-de-tronos-locuradesatada-fenomeno-mundial-brl.html.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. "The Weka Data Mining Software: An Update." *ACM SIGKDD Explorations Newsletter* 11, no. 1: 10–8.

Kilgarriff, Adam and Irene Renau. 2013. "EsTenTen, a Vast Web Corpus of Peninsular and American Spanish." *Procedia. Social and Behavioral Sciences* 95: 12–9.

Kugele, Kordula. 2010. "Analysis of Women's Participation in High-Technology Patenting." In *Innovating Women: Contributions to Technological Advancement*, edited by Pooran Wynarczyk and Susan Marlow, vol. 1, 123–51. Bingley, UK: Emerald.

Larivière, Vincent, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R. Sugimoto. 2013. "Bibliometrics: Global Gender Disparities in Science." *Nature* 504, no. 7479: 211–3.

Lax Martínez, Gema, Julio Raffo, and Kaori Saito. 2016. "Identifying the Gender of PC Inventors." *Economic Research Working Paper Nr. 33*. Geneva: World Intellectual Property Organization.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Maron, Melvin Earl. 1961. "Automatic Indexing: An Experimental Inquiry." *Journal of the ACM* 8, no. 3: 404–17.

Maslowska, Ewa, Edith G. Smit, and Bas van den Putte. 2016. "It is All in the Name: A Study of Consumers' Responses to Personalized Communication." *Journal of Interactive Advertising* 16, no. 1: 74–85.

Mosteller, Frederik, and David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist Papers*. Massachusetts: Addison-Wesley.

Motschenbacher, Heiko. 2020. "Corpus Linguistic Onomastics: A Plea for a Corpus-Based Investigation of Names." *Names* 68, no. 2: 88–103.

Naldi, Fluvio and Ilaria Vannini Parenti. 2002. "Scientific and Technological Performance by Gender." In *A Feasibility Study on Patent and Bibliometric Indicators*, edited by Henk F. Moed, Wolfgang Glänzel, and Ulrich Schmoch, 299-314. Luxembourg: European Union.

Naldi, Fulvio, Daniela Luzi, Adriana Valente, and Ilaria Vannini Parenti. 2005. "Scientific and Technological Performance by Gender." *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems*, edited by Henk F. Moed, Wolfgang Glänzel, and Ulrich Schmoch, 299–314. New York: Springer-Verlag.

National Women's Business Council. 2012. *Intellectual Property and Women Entrepreneurs: Quantitative Analysis*. Washington DC: National Women's Business Council.

Parada, Maryann. 2016. "Ethnolinguistic and Gender Aspects of Latino Naming in Chicago: Exploring Regional Variation." *Names* 64, no. 1: 19–35.

Park, Seong-Bae, and Hee-Geun Yoon. 2007. "Determining the Gender of Korean Names for Pronoun Generation." *International Journal of Computer Science and Engineering* 1, no. 4: 226–30.

Platt, John C. 1998. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. Washington DC: Microsoft Research.

Quinlan, J. Ross. 1993. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Sabet, Peyman G., and Grace Zhang. 2020. "First Names in Social and Ethnic Contexts: A Socio- Onomastic Approach." *Language & Communication* 70: 1–12.

Sugimoto, Cassidy R., Chaoqun Ni, Jevin D. West, and Vincent Larivière. 2015. "The Academic Advantage: Gender Disparities in Patenting." *PLoS One* 10, no. 5: e0128000.

Tang, Cong, Keith Ross, Nitesh Saxena, and Ruichuan Chen. 2011. "What's in a Name: A Study of Names, Gender Inference, and Gender Behavior in Facebook." *Database Systems for Advanced Applications*, edited by Jianliang Xu, Ge Yu, Shuigeng Zhou, and Rainer Unland, 344–56. Luxembourg: Springer.

Tripathi, Anshuman, and Manaal Faruqui. 2011. "Gender Prediction of Indian Names." Proceedings of the 2011 IEEE Students' Technology Symposium, 137–41. Kharagpur: IEEE.

Vapnik, Vladimir N. 1998. *Statistical Learning Theory*. New York: Wiley-Interscience.

Yoon, Hee-Geun, Seong-Bae Park, Yong-Jin Han, and Sang-Jo Lee. 2008. "Determining Gender of Korean Names with Context." *ALPIT 2008. Proceedings of the Seventh International Conference on Advanced Language Processing and Web Information Technology*, edited by Maosong Sun, Cheol Young Ock, Jeong Yong Byun, Yu De Bi, and Hong Fei Lin, 121–6. Los Alamitos, CA: IEEE.

Notes on Contributors

Rogelio Nazar is a professor of linguistics in the Institute of Literature and Language Sciences of the Pontifical Catholic University of Valparaiso, Chile. He specializes in computational linguistics and currently leads the research group Tecling.com, dedicated to the development of linguistic technologies.

Irene Renau is a professor of linguistics in the Institute of Literature and Language Sciences of the Pontifical Catholic University of Valparaiso. Her main research interest is lexical semantics and the syntax-lexicon interface. She is a member of the Chilean Language Academy.

Nicolás Acosta is a student of linguistics at the National University of Cuyo in Mendoza, Argentina. He is also an active member of Tecling.com, where he works as a software developer.

Hernán Robledo is a PhD student at the Pontifical Catholic University of Valparaiso, where he also teaches undergraduate courses in linguistics.

Maha Soliman is a PhD student at the Pontifical Catholic University of Valparaiso, where she also teaches undergraduate courses in linguistics.

Sofía Zamora is a PhD student at the Pontifical Catholic University of Valparaiso, where she also teaches undergraduate courses in linguistics.

Correspondence to: Dr. Rogelio Nazar, Instituto de Literatura y Ciencias del Lenguaje Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile. Email: rogelio.nazar@pucv.cl