Name segmentation using hidden Markov models and its application in record linkage

Segmentação de nomes por meio de modelos escondidos de Markov e sua aplicação na vinculação de registros

Segmentación de nombres a través de los modelos ocultos de Markov y su aplicación en la vinculación de registros

Rita de Cassia Braga Gonçalves [1]
Sergio Miranda Freire [1]

[1] *Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brasil.*

Correspondence
R. C. B. Gonçalves
Universidade do Estado do Rio de Janeiro.
Rua Barão de Lucena 135, apto. 304, Rio de Janeiro, RJ 22260-020, Brasil.
rb.braga.g@gmail.com

Abstract

This study aimed to evaluate the use of hidden Markov models (HMM) for the segmentation of person names and its influence on record linkage. A HMM was applied to the segmentation of patient's and mother's names in the databases of the Mortality Information System (SIM), Information Subsystem for High Complexity Procedures (APAC), and Hospital Information System (AIH). A sample of 200 patients from each database was segmented via HMM, and the results were compared to those from segmentation by the authors. The APAC-SIM and APAC-AIH databases were linked using three different segmentation strategies, one of which used HMM. Conformity of segmentation via HMM varied from 90.5% to 92.5%. The different segmentation strategies yielded similar results in the record linkage process. This study suggests that segmentation of Brazilian names via HMM is no more effective than traditional segmentation approaches in the linkage process.

Markov Chains; Information Systems; Database

Resumo

Este estudo visa avaliar a utilização dos modelos escondidos de Markov (HMM) para a segmentação de nomes de pessoas e sua influência na vinculação de registros. Um modelo HMM foi aplicado à segmentação dos nomes do paciente e da mãe nas bases do Sistema de Informações sobre Mortalidade (SIM), Subsistema de Informação de Procedimentos de Alta Complexidade (APAC), e Sistema de Informação Hospitalar (AIH). Uma amostra de 200 pacientes de cada uma das bases foi segmentada via HMM e o resultado foi comparado com a realizada pelos autores. As bases APAC-SIM e APAC-AIH foram vinculadas usando-se três diferentes estratégias de segmentação dos nomes, sendo uma delas via HMM. A conformidade da segmentação via HMM variou de 90,5% a 92,5%. As estratégias de segmentação produziram resultados similares na vinculação de registros. Este estudo sugere que a segmentação de nomes brasileiros por meio do HMM não é mais eficaz no processo de vinculação que segmentações tradicionais.

Cadeias de Markov; Sistemas de Informação; Base de Dados

http://dx.doi.org/10.1590/0102-311X00191313

Introduction

A major challenge faced by organizations is the integration of their information systems. Many institutions still work with the so-called "standalone" systems, that is, systems that work independently of others. However, eventually it is necessary arises to integrate these systems in order to improve processes or generate strategic decision-making information.

For example, the Brazilian Unified National Health System (SUS) has various health information systems developed independently and at different times. It is thus very difficult to link data from one system to the others. Integration of databases from these systems is useful to support health planning and develop new epidemiological indicators on the population's health situation, in addition to the indicators already produced by the individual databases [1]. However, this integration is limited by the difficulty in using deterministic means to determine which records belong to the same entity in the respective databases. Various approaches are used to perform database integration in such scenarios, and this is an active field of research [2]. Possibly the most widely used technique for systems integration is probabilistic linkage, especially that proposed by Fellegi & Sunter [3]. Some preliminary stages are necessary in the record linkage process: data cleaning and standardization, and blocking.

The cleaning and standardization stage involves preparation of data fields, seeking to minimize errors during the blocking and record matching process. Due to the low quality of original data completion, this stage is extremely important, contributing greatly to the efficiency of the process. Another important component of standardization is segmentation (separation) of the name into its constituent parts. The objective is to increase (insofar as possible) the likelihood that a given individual will be truly identified.

There are various probabilistic record linkage software programs that include a segmentation stage. In Brazil, Camargo Jr. & Coeli [4] developed a free software for probabilistic record linkage (*Reclink*) that includes a stage involving separation of the person's name into first name, surname, initials of the middle names, and suffixes (Filho, Junior, etc.). *Reclink* has made an important contribution to the use of record linkage in the health field in Brazil. However, this does not rule out the study of other alternatives for name segmentation that could make the linkage process more efficient than that proposed by *Reclink*. For example, the software developed by the Australian National University – *Freely Extensible Biomedical Record Linkage* – Febrl [5] has tools for segmentation of names and addresses via hidden Markov models (HMM) [6]. The latter authors applied HMM to English-language names [7].

The objective of the current study was to apply HMM to the segmentation of Brazilian names and verify whether use of the parts of the name thus obtained in the record linkage process is more efficient than the traditional name segmentation methods. The basic assumption is that the use of initials from middle names leads to loss of information and that use of all complete parts of the name would result in greater efficiency in the linkage process.

Materials and methods

This section is divided into four subsections for greater clarity in the presentation: databases used, name segmentation process, evaluation of segmentation, and evaluation of the influence of segmentation on the record linkage process.

Databases used

The study used the databases from the Mortality Information System (SIM) [8] referring to records from Rio de Janeiro State from 1999 to 2004, the Subsystem for Authorization of High Complexity Procedures (APAC) [9], and the Subsystem for Authorization of Hospital Admissions, or Hospital Information System (AIH) [10], all referring to Rio de Janeiro State from 2000 to 2004. For the purposes of this study, only the patient's name and the patient's mother's name were considered. The APAC and SIM databases were used to construct the ancillary tables and generate the HMM, as explained in the next section. The AIH database was used to apply one of the previously obtained HMM without the need to generate new ancillary tables or alter the existing tables.

Name segmentation process

The methodology used for name segmentation consists of eight phases: data cleaning, standardization of the form, name standardization, name segmentation, creation of the initial HMM, training, and refinement.

The data cleaning phase identified records that were invalid for linkage and performed corrections in the name field, preparing it for the subsequent standardization phases.

Standardization of the form included some corrections and/or substitutions of some spelling variations according to a standard established for representing the name's form: capitalization of the letters; removal of accent marks; removal of spaces at the beginning and end of the name; removal of double spaces; removal of prepositions; and removal of punctuation marks.

The name standardization phase created "dictionary" tables. These tables consisted of two fields, current_name and correct_name, and functioned as follows: when a term from the name was found in the current_name table, the term was corrected according to the correct_name field. For example, this process can replace all the variations for the surname "GONCALVES", such as "GONCAVES", "GONEALVES", "GONCAOLVES". Three such tables were created, for given names (dic_name), surnames (dic_surname), and suffixes (dic_suffix).

The name segmentation phase was subdivided into two stages. In the first stage, names were separated into five distinct fields. The target databases (APAC and SIM) had 99% of their records with names containing fewer than 6 parts. For the remaining 1% of records, whose names contained more than five parts, the criterion for eliminating parts of the name was based on the hypothesis that the outermost parts of the name are the most important for the linkage process. *Reclink* adopts a similar hypothesis, since it only uses the initials from middle names. Thus, the following adjustment was performed for names with more than 5 parts: (a) names with six parts: the 4th part of the name was eliminated; (b) names with seven parts: the 4th and 5th parts of the name were eliminated; (c) names with eight parts: the 4th, 5th, and 6th parts of the name were eliminated; and (d) names with nine or more parts: the first three and the last two parts of the name were maintained.

Based on separation of the name into its constituent parts and using the "dictionary" tables, each part was identified with a qualifier (symbol) corresponding to its meaning. For example: the qualifier for João would be "male name - MN", because the word "João" is found in the dictionary table for given names. The qualifier for Silva would be "surname – SN" because the word "Silva" is found in the dictionary table for surnames. Applying similar reasoning, the qualifier for "Junior" would be "suffix – SU".

The output of this stage was the name (string field) separated into its parts with one or more qualifiers identifying the part belonging to one (or more) of the "dictionary" tables. The qualifiers may thus become incorrect.

To understand the segmentation, consider a simple example of a name:
Conceicao Maria Lucena → ['Conceicao', 'Maria', 'Lucena']

According to the "dictionary" tables, the parts would be classified with the following qualifiers:
['Conceicao', 'Maria', 'Lucena']
['FN' 'FN' 'SN']
['SN' 'FN' 'SN']
where: FN = female name and SN = surname.

The problem of selecting the most likely sequence was solved with a probabilistic model called the HMM [7]. The main underlying idea in the model is that there are various phenomena whose outputs depend on factors that are not directly observable (i.e., they are hidden), but that can be inferred from these outputs. The model's use allows making a statistical distinction between these hidden factors, separating them into different states in a Markov chain [11].

A HMM consists of: (1) a set of hidden states S; (2) a probability of transition $P[s'|s]$ between hidden states s *e* $s' \in S$; (3) a set of symbols (observations) T emitted by the hidden states; and (4) a probability distribution of symbol emissions for each hidden state. The annotation $P[t|s]$ gives the probability of emission of symbol $t \in T$ for the hidden state $s \in S$ [12].

In the above-mentioned example, one can assume that a hidden Markov model for the name field would have the following states: first name, second name, first surname, second surname, and third surname. These would be the hidden states of the above-mentioned set S. Each identification symbol is assumed to be emitted by a hidden state. Thus, the sequences of states could be the following:
Start → First given name [FN] → Second given name [FN] → First surname
Start → First surname [SN] → First given name [FN] → Second surname

Intuitively, the first sequence would be more probable than the second, indicating that this sequence of hidden states is more consistent with the sequence of symbols. This probability is calculated using the *Viterbi* algorithm [6], which produces the sequence of hidden states with the highest probability of having emitted each sequence of input symbols.

The HMM was defined as follows. The hidden states are: given name 1, given name 2, surname 1, surname 2, and surname 3. The symbols are:
FN: female name (found in the dic_name table indicated as female name);
MN: male name (found in the dic_name table indicated as male name);
SN: surname (found in the dic_surname table);
SU: suffix (found in the dic_suffix table);
UN: unknown (name not found in any table);
IL: initial letter (only a letter corresponding to an abbreviation of the name).

In the next phase, creation of the initial hidden Markov model, a thousand random records were selected from the APAC and SIM databases and the respective sequences of identification symbols were generated. The sequences were used to calculate the probabilities of state transitions, probability distributions of symbol emissions for each hidden state, and the initial state's vector, thereby defining an initial hidden Markov model for each database.

The training and refinement phases aim to achieve the best fit of the initial model to the real data. The sequence of observations used to make this fit is called a "training sequence", since it is used to train the HMM. For each phase, another random sequence of a thousand records was selected from APAC and SIM, generating the corresponding identification symbols. The *Baum-Welch* algorithm [13] was used to adjust the initial model's parameters. The algorithm is a method of iterative re-estimation which generates (for each new model) a sequence of observations with higher probability than the previous model. The new model was estimated, starting from the initial model and the training sequence, using JAHMM (Jahmm hidden Markov model) [14], which implements open source HMM algorithms in Java language. Iterations of the model were done, and the *Kullback-Leibler* divergence [14] between the two models was calculated; the iterations were interrupted when divergence between two consecutive models dropped below 10^{-5}.

All the tables created in these stages, as well as the algorithms used in the process, can be obtained by consulting the authors.

Evaluation of name segmentation via HMM

To evaluate the quality of the segmentation generated by the hidden Markov model, 200 records were selected randomly from the APAC and SIM databases, and "dictionary" tables were used to generate the corresponding identification symbols. The *Viterbi* algorithm [12] was used to determine the best sequence of states run by the model for the sequences of observations generated in the previous stage. With the estimated model and the sequences of observations for each name, the JAHMM library was used to determine the sequence of hidden states for each name.

Next, the hidden Markov model generated for the names in the APAC database was also used to segment a random sample of 200 records from the AIH database, using the same dictionary tables generated from APAC and SIM.

One of the authors was in charge of creating the tables and training the hidden Markov model. With this author as the reference, the hidden Markov model's conformity was evaluated by the proportion of hits in the sequence of states generated by the names of the test samples. This study adopted the terminology proposed by Müller & Buttner [15], which defines conformity as the agreement between two observations when one is taken as the reference or standard, and consistency as the agreement between two observations when neither can be taken as the reference.

The authors evaluated the sequences of states generated for the sequences of observations in order to assess the consistency between two independent reviewers, measured by the kappa coefficient [16]. The cells in the 2 x 2 table for estimating the kappa coefficient indicate the number of times in which the sequences of hidden states generated by the hidden Markov model were classified, respectively, as: correct by both reviewers; correct by reviewer A, but incorrect by reviewer B; correct by reviewer B, but incorrect by reviewer A; and incorrect by both reviewers. The confidence intervals for the measurement of conformity were calculated with the OpenEpi software (version 3.0.1) (Dean AG, Sullivan KM, Soe MM. OpenEpi: Open Source Epidemiologic Statistics for Public Health, http://www.OpenEpi.com) and those for the kappa coefficient were calculated as presented by Fleiss [17].

Evaluation of the influence of segmentation on the record linkage process

In order to evaluate the application of name segmentation via HMM in record linkage of SIM and APAC, 20 thousand records were randomly selected from each of the respective databases. Linkage was done with a software developed by the Medical Information Technology team at the Rio de Janeiro State University (UERJ). The fields selected for record linkage were: patient's full name, mother's full name, date of birth, and sex.

Three linkage processes were performed, each with a different name segmentation strategy. The first was that used by the *Reclink* software [4], excluding the name's suffix. The second segmentation consisted of separating the name into a maximum of five parts before applying the HMM. The third segmentation was the name segmentation resulting from the application of the hidden Markov model to the parts of the name obtained from the second strategy, identifying whether each part of the name was given name 1 (GN1), given name 2 (GN2), surname 1 (LN1), surname 2 (LN2), and so on. As an example, the name João Batista Souza da Silva would be broken into the following parts:

Second alternative: P1 – JOAO; P2: BATISTA; P3: SOUZA; P5: SILVA
Third alternative: GN1 – JOAO; GN2: BATISTA; LN1: SOUZA; LN2: SILVA

In the second alternative, the last part of the name was always placed in P5, regardless of the number of parts in the name.

Blocking was performed in two stages: the first part of the patient's name and last part of the mother's name were coded phonetically using an adaptation of the *Soundex* algorithm, as implemented in the *Apache Commons Project* [18], having introduced the modifications proposed by Coeli & Camargo Jr. [19].

In a previous study [1], record linkage was performed in the SIM and APAC databases. Taking this previous study as the basis for estimating the parameters m_i (probability of agreement between values for variable i, assuming that the pair of compared records is true) for variable i, the following steps were performed: (1) 248 pairs of records sampled from the SIM and APAC tables were identified that were considered true in the previous study [1]; (2) for variable, m_i was estimated as the amount of these pairs for which the variable's values agreed in the two records for each pair, divided by the total number of true pairs.

To estimate the parameters u_i (probability of agreement between values for variable i, assuming that the pair of compared records is false) for each variable i, the following steps were performed: (1) 100 random records from the sampled APAC table were paired with 100 random records from the sampled SIM table, for a total of 10,000 pairs of records; (2) for each variable, u_i was estimated as the amount of these pairs for which the values of the variable agreed in the two records for each pair, divided by the total number of pairs (10,000).

The final stage in the record linkage process was to define the cutoff point. For each strategy, cutoff points were established through manual inspection by two reviewers. The pairs with scores above the cutoffs points were classified by consensus by the authors as false or true. Taking the pairs classified as false or true as a gold standard, it was possible to evaluate the efficiency of the linkage according to the following metrics [20]: (a) recall rate, defined as the proportion of real pairs of existing records that were classified as such in the linkage process and (b) precision rate, defined as the proportion of pairs of records identified in the linkage process that were real pairs.

An analogous linkage process was performed in the 20,000 records from APAC as obtained above with 20,000 thousand random records from AIH, using patient's name, date of birth, and sex as the variables.

The research project was approved by the Ethics Research Committee of the Pedro Ernesto University Hospital (CEP/HUPE – CAAe:0153.0.228.000-07).

Results

HMM were generated for patient's name and mother's name in the SIM and APAC databases. Figure 1 shows the model for patient's name in the SIM database. The model shows the hidden states and probabilities of transition between these states. Thus, using the annotation for the HMM, P[Surname 1/Given name 1] = 0.655, and so on. The model reflects the composition of Brazilian names that begin with a given name followed by a surname, or less frequently with a second given name (P[Given name 2/Given name 1] = 0.345). Occasionally the full name may include three given names (P[Given name 3/Given name 2] = 0.030). In the database, if the state of the given name 2 appears twice, the second occurrence is stored as a third given name. Starting with the first surname, all the subsequent parts of the name are classified as surnames.

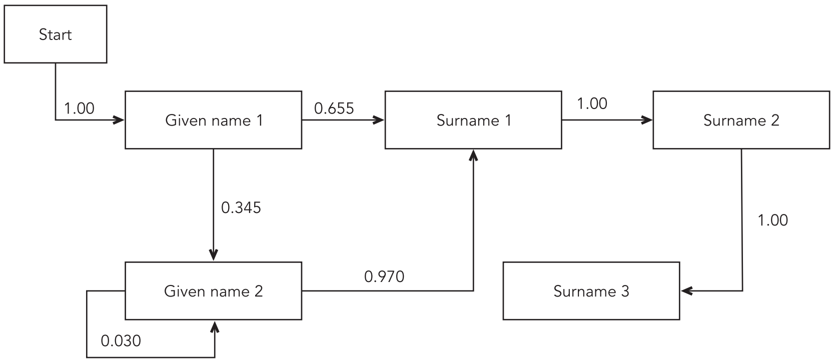
Figure 2 shows the model for patient's name in the APAC database. This model is quite similar to that of Figure 1, and the same is true comparing the model for mother's name in the APAC and SIM databases.

Table 1 shows the probability distributions for the emission of observations for each hidden state in the same models for Figures 1 and 2. Thus, in the state of given name 1 in the SIM database, the probability of observing the symbol FN (female name) is 0.353 (P[FN/Given name 1] = 0.353), and so on. Note the low probability of observing the symbols UN (unknown) and SU (suffix) in any state.

Table 2 shows the conformity and consistency values for the name segmentations in each of the databases. The conformity values for the segmentations generated by the hidden Markov models were greater than 90%, taking as the reference the observer that conducted the model training. Inter-observer consistency, measured by kappa, was substantial, considering the classification proposed by Landis & Koch [22], although for patient's name in the APAC and SIM samples and for mother's name in the SIM samples, the confidence interval suggested moderate consistency. The patient's name in the SIM sample showed the lowest inter-observer consistency (k = 0.64). Of the 200 names in this sample, the observers disagreed on the segmentation in 15.

Figure 1

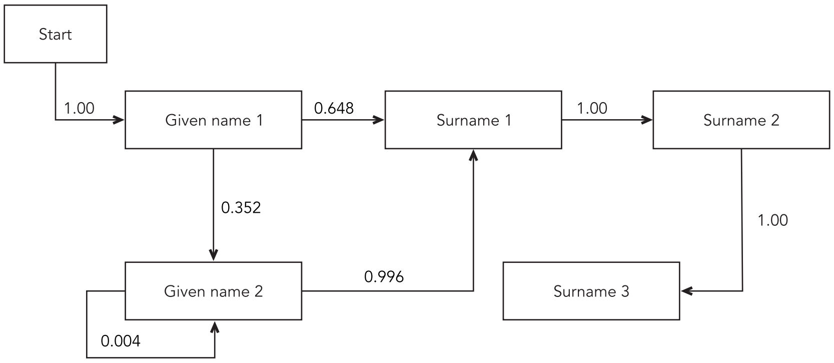
HMM for segmentation of person's name in the SIM database. Rectangles represent the hidden states, and the adjoining numbers or those over the arrows connecting the states represent the probability of transition between the states.



HMM: hidden Markov model; SIM: Mortality Information System (Brazil).

Figure 2

HMM for segmentation of patient's name in the APAC databases. Rectangles represent the hidden states, and the adjoining numbers or those over the arrows connecting the states represent the probability of transition between the states.



APAC: Information Subsystem for High Complexity Procedures; HMM: hidden Markov model.

Disagreement occurred mainly due to parts of the name that could be classified as either given names or surnames, depending on the observer's point of view.

Table 3 shows the recall and precision metrics for the APAC-SIM and APAC-AIH record linkages, using three different segmentation strategies. The segmentation produced by HMM had the lowest impact on both linkage processes, although the differences between the three strategies were not statistically significant.

Discussion

Record linkage is an active area of research [2]. Even considering that more widespread use of the national health card number in health systems

Table 1

Matrices for emission of symbols (observations) for names in the SIM/APAC databases. The probabilities of emission are separated by a slash, with the first probability for the name in SIM and the second for the name in APAC.

Hidden states	**Symbols (observations)**					
	FN	**MN**	**SN**	**IL**	**UN**	**SU**
Given name 1	0.353/0.706	0.515/0.191	--	--	0.122/0.103	--
Given name 2	0.377/0.799	0.623/0.201	--	--	--	--
Surname 1	--	0.005/0.022	0.875/0.797	0.013/0.060	0.104/0.120	0.003/0.001
Surname 2	--/0.005	0.020/0.038	0.858/0.804	0.016/0.048	0.073/0.105	0.018/0.005
Surname 3	--/0.016	0.031/0.055	0.732/0.537	0.015/0.112	0.139/0.262	0.067/0.034

Two hyphens (--) indicate that the probability of emission is nil.

APAC: Information Subsystem for High Complexity Procedures; FN: female name; IL: initial letter; MN: male name; SIM: Mortality Information System (Brazil); SN: surname; SU: suffix; UN: unknown.

Table 2

Evaluation of segmentation via HMM. Measures of conformity and consistency, with respective 95% confidence intervals (95%CI), are shown for segmentation of person's name and mother's name in the APAC and SIM databases, and for the patient's name in the AIH database.

Variable	**Conformity**		**Consistency (kappa)**	
	Value (%)	**95%CI**	**Value**	**95%CI**
Name, APAC	92.5	87.7-95.6	0.67	0.50-0.84
Mother's name, APAC	90.5	85.4-94.0	0.76	0.63-0.90
Name, SIM	91.5	86.5-94.8	0.64	0.48-0.81
Mother's name, SIM	92.0	87.1-95.2	0.69	0.52-0.86
Name, AIH	91.5	86.5-94.8	0.88	0.77-0.99

AIH: Hospital Information System; APAC: Information Subsystem for High Complexity Procedures; HMM: hidden Markov models; SIM: Mortality Information System (Brazil).

Table 3

Influence of segmentation on record linkage. Recall and precision measurements are shown for linkage between APAC and SIM, and APAC and AIH, using three different name segmentation strategies in each database.

Linkage/Type of segmentation	**Recall**		**Precision**	
	Value (%)	**95%CI**	**Value**	**95%CI**
APAC-SIM				
Strategy 1	98.42	95.73-99.49	96.51	93.26-98.29
Strategy 2	98.42	95.73-99.49	94.68	91.03-96.96
Strategy 3	96.44	93.13-98.25	94.94	91.30-97.17
APAC-AIH				
Strategy 1	94.96	88.89-97.93	100.00	95.90-100.00
Strategy 2	95.80	89.98-98.44	99.13	94.55-99.95
Strategy 3	94.96	88.89-97.93	91.13	84.33-95.27

AIH: Hospital Information System; APAC: Information Subsystem for High Complexity Procedures; HMM: hidden Markov models; 95%CI: 95% confidence interval; SIM: Mortality Information System (Brazil).

could facilitate integration of these systems, record linkage using non-deterministic means will still be used to integrate the older records in these systems, as well as in systems that do not use a unique identifier. This study investigated the application of hidden Markov model as a mechanism for segmentation of the name field within the cleaning and standardization stage for record linkage.

The person's name, essential for determination of the link between a pair of records, is a variable whose treatment must be investigated in order to obtain the most efficient strategies for record linkage. In Brazil, *Reclink* [4] has been widely used to perform record linkage in the health area. Thus, the proposal of *Reclink* for segmentation of names has been the most widely used. Queiroz et al. [22] used another segmentation strategy: first name, surname, and all the internal parts of the name considered as one, after eliminating prepositions.

Martinhago [23] used Febrl [5] to perform name segmentation in a data set in the Library System of the Federal University of Paraná, Brazil. The model did not include prepositions, the word "e" (Portuguese for "and"), or suffixes as states. Considering only the transitions from given names to given names and from given names to surnames, one observes a great similarity between the probabilities found in the model and in the models obtained in the current study. However, neither Martinhago [23] nor Queiroz et al. [22] evaluated the efficiency of their segmentation proposals in record linkage, as compared to other segmentation strategies.

Name segmentation via HMM showed excellent agreement with the segmentation generated by the observer that trained the models. In the majority of cases, the disagreement related to compound given names, where the model classified the second given name as a surname. This was also the most frequent disagreement when the two observers conducted the name segmentation. The size of the samples should be increased in order to improve the precision estimated conformity and consistency.

During the database preparation phase, a major effort was made to create "dictionary" tables. These tables served to assist the data cleaning and standardization. According to the findings, 57% of the records in both the SIM and APAC databases had some change made to the patient's name, based on the patterns established in these tables. For example, 30 different written forms were recorded for the name "Conceicao". These tables are available to the community, for the time being after prior consultation with the study's authors. The tables can be useful in record linkage applications, regardless of the type of name segmentation used. The current study used these tables without alterations to perform linkage of the APAC and AIH databases. The progressive inclusion of more records in these tables as new names or variations are identified would obviously improve the precision of segmentations and reduce the number of parts of the name that are unknown to the models.

The segmentation strategies evaluated here presented very similar results in the linkage process, while strategy 1 (based on *Reclink*) performed better than segmentation via HMM and similarly to the simple splitting of names, although the differences were not statistically significant. A possible explanation for these findings is that the estimates for m_i are lower and that those for u_i are higher for parts of the name that occur less frequently, e.g., the second given name or second surname, implying that agreement between these parts bears less weight in the pair's score.

Name segmentation via HMM is a painstaking computational process when compared to other simpler forms of segmentation. The current study suggests that segmentation produced by HMM, when applied to record linkage, does not produce better results than traditional segmentation methods. Although this was a "negative" finding, the authors believe in the importance of publishing the results in order to avoid publication bias in studies with findings that contradict the authors' expectations. This kind of bias has been detected in different fields of knowledge [24,25] and is still present in the medical literature [26].

However, caution is recommended in generalizing the study's findings. The study's main limitations are: sizes of the samples, both for evaluating name segmentation and for the database linkage, a fixed linkage scenario in terms of the variables chosen for comparison, comparators used to compare the variables, databases used, and linkage technique. This type of study should be reproduced in other scenarios, for example: different databases; other name segmentation strategies [22]; other linkage techniques [2]; different weighting strategies for the variables, for example weighting based on the name's frequency [22,27,28]; other string comparators [29]; and use of address segmentation. It would be interesting to draw on publically available databases in order for the various research groups that work with database integration to have a common platform for conducting simulations with their proposals.

Resumen

Este estudio tiene como objetivo evaluar el uso de los modelos ocultos de Markov (HMM) para la segmentación de nombres y de su influencia en la vinculación de registros médicos. Los modelos HMM se aplicaron a la segmentación de los nombres del paciente y de la madre en las bases del Sistema de Información sobre Mortalidad (SIM), Subsistema de Información para los procedimientos de alta complejidad (APAC), y Sistema de Información Hospitalaria. Una muestra de 200 pacientes de cada base fue segmentada por HMM y el resultado se comparó con la obtenida por los autores. Las bases APAC-SIM y APAC-AIH se vincularon con 3 diferentes estrategias de segmentación, siendo una de ellas por HMM. La conformidad de la segmentación por HMM varió de 90,5% a 92,5%. Las estrategias dieron resultados similares en la vinculación. Este estudio sugiere que la segmentación de nombres brasileños por HMM no es más eficaz en el proceso de vinculación que la segmentación tradicional.

Cadenas de Markov; Sistemas de Información;
Base de Datos

Contributors

R. C. B. Gonçalves participated in the conceptualization, modeling, implementation, and writing of the article. S. M. Freire participated in the conceptualization and writing of the article.

Acknowledgments

The authors wish to thank the anonymous reviewers who contributed to improvement of the article's wording and presentation.

References

1. Sousa RC. Desenvolvimento de um armazém de dados a partir da integração de sistemas de informação em saúde para apoiar a gestão da assistência oncológica [Doctoral Dissertation]. Rio de Janeiro: Universidade do Estado do Rio de Janeiro; 2012.
2. Christen P. Data matching. Concepts and techniques for record linkage, entity resolution, and duplicate detection. Berlin: Springer-Verlag; 2012.
3. Fellegi IP, Sunter A. A theory of record linkage. J Am Stat Assoc 1969; 64:1183-210.
4. Camargo Jr. KR, Coeli CM. *Reclink*: aplicativo pa-ra o relacionamento de bases de dados, implementando o método *probabilistic record linkage*. Cad Saúde Pública 2000; 16:439-47.
5. Christen P. Febrl – a freely available record linkage system with a graphical user interface. In: Warren JR, Yu P, Yearwood J, Patrick JD, editors. Proceedings of the Second Australasian Workshop on Health Data and Knowledge Management, 2008. http://crpit.com/confpapers/CRPITV80Christen.pdf (accessed on 10/Oct/2011).
6. Rabiner L, Juang B. An introduction to Hidden Markov Models. IEEE Acoustic, Speech, and Signal Processing Magazine 1986; 3:4-16.
7. Churches T, Christen KL, Zhu JX. Preparation of name and address data for record linkage using hidden Markov models. BMC Med Inform Decis Mak 2002; 2(9). http://www.biomedcentral.com/content/pdf/1472-6947-2-9.pdf.

8. Ministério da Saúde. Manual de instrução do Sistema de Informações sobre Mortalidade. http://bvsms.saude.gov.br/bvs/publicacoes/sis_mortalidade.pdf (accessed on 02/Apr/2014).
9. Ministério da Saúde. APAC – Autorização de Procedimento Ambulatorial: manual de operação do sistema. ftp://arpoador.datasus.gov.br/siasus/documentos/Manual_Operacional_APAC_v1.pdf (accessed on 02/Apr/2014).
10. Ministério da Saúde. Manual técnico do sistema de informação hospitalar. http://bvsms.saude.gov.br/bvs/publicacoes/07_0066_M.pdf (accessed on 02/Apr/2014).
11. Souza-e-Silva EG. Investimentos no mercado de petróleo: uma abordagem utilizando modelos de Markov ocultos [Masters Thesis]. Rio de Janeiro: Instituto Alberto Luiz Coimbra de Pós-graduação e Pesquisa de Engenharia, Universidade Federal do Rio de Janeiro; 2009.
12. Silva EFA, Barros FA, Prudêncio RBC. Uma abordagem de aprendizagem híbrida para extração de informação em textos semi-estruturados. In: Anais do XXV Congresso da Sociedade Brasileira de Computação, v. 1. http://www.lbd.dcc.ufmg.br/colecoes/enia/2005/047.pdf (accessed on 10/Oct/2013).
13. Baum LE, Petrie T, Soules G, Weiss NA. Maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Annals of Mathematical Statistics 1970; 41:164-71.
14. Francois JM. Jahmm-hidden Markov model (HMM): an implementation in Java. http://jahmm.googlecode.com (accessed on 09/Sep/2013).
15. Müller R, Buttner P. A critical discussion of intraclass correlation coefficients. Stat Med 1994; 13:2465-76.
16. Cohen JA. Coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 20:37-46.
17. Fleiss JL. Statistical methods for rates and proportions. 2nd Ed. New York: John Wiley & Sons; 1981.
18. Apache Commons Project. Implementations of common encoders and decoders. http://commons.apache.org/codec (accessed on 20/Mar/2012).
19. Coeli CM, Camargo Jr. KR. Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. Rev Bras Epidemiol 2002; 5:185-96.
20. Herzog T, Scheuren F, Winkler W. Data quality and record linkage techniques. New York: Springer Science; 2007.
21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159-74.
22. Queiroz OV, Guerra Jr. AA, Machado CJ, Andrade ELG, Meira Jr. W, Acúrcio FA, et al. A construção da base nacional de dados em terapia renal substitutiva (TRS) centrada no indivíduo: relacionamento dos registros de óbitos pelo subsistema de autorização de procedimentos de alta complexidade (APAC/SAI/SUS) e pelo sistema de informações de mortalidade (SIM) – Brasil, 2000-2004. Epidemiol Serv Saúde 2009; 18:107-20.
23. Martinhago AZ. Customização em ambientes de qualidade de dados [Masters Thesis]. Curitiba: Universidade Federal do Paraná; 2006.
24. Rosenthal R. Combining probabilities and the file drawer problem. Evaluation in Education 1980; 4:18-21.
25. Egger M, Smith GD. Bias in location and selection of studies. BMJ 1998; 316:61-6.
26. Dwan K, Gamble C, Williamson PR, Kirkham JJ; for the Reporting Bias Group. Systematic review of the empirical evidence of study publication bias and outcome reporting bias – an updated review. PLoS One 2013; 8:e66844.
27. Yancey WE. Frequency-dependent probability measures for record linkage. Bureau of the Census Statistical Research Division. Research Report Series; 2000. http://www.census.gov/srd/papers/pdf/rr2000-07.pdf (accessed on 10/Nov/2013).
28. Gill L. Methods for automatic record matching and linking and their use in national statistics. National Statistics Methodology series 25; 2001. http://www.ons.gov.uk/ons/guide-method/method-quality/specific/gss-methodology-series/gss-methodology-series--25--methods-for-automatic-record-matching-and-linkage-and-their-use-in-national-statistics.pdf (accessed on 02/Apr/2014).
29. Freire SM, Gonçalves RCB, Bandarra AC, Villela MGT, Meire A, Cabral MDB, et al. Análise da efetividade de comparadores de strings para discriminar pares verdadeiros de pares falsos no relacionamento de registro. In: Ziviani A, Ito M, Nedel LP editors. Anais do IX Workshop de Informática Médica. XXIX Congresso da Sociedade Brasileira de Computação – IX Workshop de Informática Médica. Bento Gonçalves: Sociedade Brasileira de Computação; 2009. p. 2119-28.

Submiited on 09/Nov/2013
Final version resubmitted on 30/Jun/2014
Approved on 22/Jul/2014