International Journal of Medical Informatics 110 (2018) 111–117

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

# Inferred joint multigram models for medical term normalization according to ICD

Alicia Pérez[b], Aitziber Atutxa[b], Arantza Casillas[a,*], Koldo Gojenola[b], Álvaro Sellart[b]

[a] Dep. Electricity and Electronics, Faculty of Science and Technology, IXA Research Group, University of the Basque Country (UPV-EHU), Spain[1]
[b] Dep. Languages and Computer Systems, Technical School of Engineering of Bilbao, IXA Research Group, University of the Basque Country (UPV-EHU), Spain

ARTICLE INFO

*Keywords:*
International Classification of Diseases
Electronic Health Records
Normalization
Finite State Models

ABSTRACT

*Background:* Electronic Health Records (EHRs) are written using spontaneous natural language. Often, terms do not match standard terminology like the one available through the International Classification of Diseases (ICD).
*Objective:* Information retrieval and exchange can be improved using standard terminology. Our aim is to render diagnostic terms written in spontaneous language in EHRs into the standard framework provided by the ICD.
*Methods:* We tackle diagnostic term normalization employing Weighted Finite-State Transducers (WFSTs). These machines learn how to translate sequences, in the case of our concern, spontaneous representations into standard representations given a set of samples. They are highly flexible and easily adaptable to terminological singularities of each different hospital and practitioner. Besides, we implemented a similarity metric to enhance spontaneous-standard term matching.
*Results:* From the 2850 spontaneous DTs randomly selected we found that only 7.71% were written in their standard form matching the ICD. This WFST-based system enabled matching spontaneous ICDs with a Mean Reciprocal Rank of 0.68, which means that, on average, the right ICD code is found between the first and second position among the normalized set of candidates. This guarantees efficient document exchange and, furthermore, information retrieval.
*Conclusion:* Medical term normalization was achieved with high performance. We found that direct matching of spontaneous terms using standard lexicons leads to unsatisfactory results while normalized hypothesis generation by means of WFST helped to overcome the gap between spontaneous and standard language.

## 1. Introduction

Electronic Health Records (EHRs) contain vast amounts of valuable data. Several services (pharmacy, documentation, etc.) exchange and consult medical information in the form of EHRs on a daily basis. The International Classification of Diseases (ICD) serves as a reference to exchange this kind of information (e.g., billing, epidemiologies or mortality) between hospitals in a country and even between countries. Regarding the economic impact of the task, Lang mentioned in [1] that the cost of coding clinical text with ICD in US is $25 billion per year.

However, the encoding of diagnoses according to ICD is a difficult, time consuming and expensive task for health services, because these records are written in a spontaneous natural language and this lack of standardization is of particular relevance for Diagnostic Term (DT) reporting. In [2] the authors presented a study on reliability of human encoding of diagnoses within the ICD and they concluded that codification might result uncertain even for human experts. Most hospitals, currently, hire expert human encoders to manually assign an ICD to each spontaneous non-standard DT.

The **challenge** arises from the high divergence between spontaneous and standard language used to retrieve and exchange information. The **aim** of this work is to tackle language normalization in the context of the diagnostic term coding according to the ICD. The following are examples of ICD codes with the standard (Std) and spontaneous (Spo) DTs found in EHRs:

- `ICD code`: `571.2`
  - `Std-DT`: “*cirrosis hepática alcohólica*”
  - `Spo-DT`: “*ci OH*”
- `ICD code`: `303.90`
  - `Std-DT`: “*otras dependencias alcohólicas y dependencias alcohólicas no especificadas*”
  - `Spo-DT`: “*etilismo*”

[*] Corresponding author at: Dep. Electricity and Electronics, Faculty of Science and Technology, UPV-EHU, Barrio Sarriena s/n, 48940 Leioa, Bizkaia, Spain.
*E-mail addresses:* alicia.perez@ehu.eus (A. Pérez), aitziber.atucha@ehu.eus (A. Atutxa), arantza.casillas@ehu.eus (A. Casillas), koldo.gojenola@ehu.eus (K. Gojenola).
[1] http://ixa.eus

https://doi.org/10.1016/j.ijmedinf.2017.12.007
Received 29 September 2017; Received in revised form 10 November 2017; Accepted 11 December 2017

Regarding the literature on automatic classification of clinical text with respect to ICD, the BioNLP workshop 2007 organized a shared task with the objective of assigning ICD-9-CM codes to radiology reports in English [3]. The best system [4] achieved an *f*-score of 76.35 and it consisted of a hybrid model that used a small set of hand-crafted rules. In [5], the authors trained machine learning models to automatically construct the rules. The main shortcoming of this task is that the number of ICD-codes is reduced compared to our setting (45 ICD codes with respect to thousands of codes in our case).

In 2017 the CLEF eHealth task [6] was devoted to the extraction of ICD-10 codes from short notes related to death certificates in French and English. The best systems obtained an *f*-score of 80.4 for French and 85.0 for English. There were approaches based on machine learning such as neural nets or classifiers, and also knowledge-based approaches. As these tasks are performed on a high number of ICD codes (over 3000 codes for French and 1300 for English), they can be comparable to our challenge, although, we focus on Spanish.

Other authors have also reported the high lexical variability found in medical records [7–9]. Indeed, Dziadek et al. [9] turned to context-sensitive spelling correction to normalize spontaneous terms with respect to SNOMED-CT. Spelling correction tends to rely upon simple regular expressions that can be modeled with WFSTs, which can also propose the alternatives and rank them.

Rule-based systems have been proved to be successful for solving normalization tasks, but they require a big effort for manually generating the rules. Bearing in mind antecedent works, we decided to focus on the development of a system that would infer, from data, transformation rules to tackle the DT normalization problem. The contribution of this work rests on the use of inferred Weighted Finite-State Transducers (WFSTs) to tackle DT normalization. This WFST will be inferred from a parallel set of samples so that it has the ability to adapt to a specific domain or a specific hospital. In addition to the DT normalization stage, we propose the use a soft-matching strategy to gain coverage. Very close to spelling correction as tackled by [9], we incorporated text similarity instead. In our case, the system was inferred from EHRs written by around 400 different practitioners, where the spontaneous DTs were manually encoded after the patient notes were extracted. It obtains the correct ICD code between the first and second positions of a ranked list. These achievements prove that our approach is able to adapt to different writing-schemes and also that it is adequate as a decision support tool, given that the hypothesis selection can be carried out from a short list.

## 2. Materials and methods

Our approach, depicted in Fig. 1, comprises two stages (marked with blue and orange rectangles respectively): (1) produce normalized alternatives to a given spontaneous string by means of WFSTs (details are given in Section 2.1); (2) apply text similarity to select the standard string that matches best the normalized alternatives (Section 2.2). The following example from the development set might help understanding the steps. The partitioner wrote *"cervicobraquialgia"* spontaneously. The human annotator assigned `723.3` as ICD. The first WFST proposition was *"sindrome cervicobraquial difuso"* and after applying text symilarity mapping, the system found that the closest dictionary DT was *"sindrome cervicobraquial (difuso)"*, with which it guessed the right ICD, in fact. Note that the practitioner used a Greek suffix (*"-algia"*) in the spontaneous DT and the dictionary DT follows the Spanish preferred form and, still, the WFST managed to bridge the gap.

### 2.1. Joint n-multigram models for text normalization

Classical *n*-gram models assume that the statistical dependencies between tokens are of a fixed length, n, along the entire sequence of tokens [10]. Joint multi-gram models were introduced in the field of automatic phonetic transcription [11]. In this work, the n-multigram models conceive a sequence of n tokens that comprise independent variable-length sub-sequences of spontaneous and standard registers (*x*-grams and *y*-grams respectively) compatible with a segmentation of the pair of sequences. That is, the segmentation is introduced as a hidden variable. Accordingly, given a compatible segmentation of length $q$ of the string $\mathbf{t} = t_1^T$, $\mathbf{s} = s_1^q$, in which each segment, $s_i = t_{1+j_{i-1}}{}^{j_i}$, has a maximum length of n tokens (that is, $|s_i| = |t_{1+j_{i-1}}{}^{j_i}| = j_i - j_{i-1} \leq n$), the n-multigram model computes the joint likelihood of the string and the segmentation assuming conditional independence between the segments as stated in (1).

$$\ell(\mathbf{t}, \mathbf{s}) = \prod_{i=0}^{q} P(s_i) \tag{1}$$

Furthermore, in order to compute the likelihood of the string $\mathbf{t}$, the sum of joint likelihoods is approached by its maximum term as in (2) where $\mathscr{S}_{(\mathbf{t},n)}$ denotes the set of all possible segmentations compatible with $\mathbf{t}$ with segments of maximum length $n$.

$$\ell(\mathbf{t}) = \sum_{\mathbf{s}\in\mathscr{S}_{(\mathbf{t},n)}} \ell(\mathbf{t}, \mathbf{s}) \tag{2}$$

$$\approx \max_{\mathbf{s}\in\mathscr{S}_{(\mathbf{t},n)}} \ell(\mathbf{t}, \mathbf{s}) \tag{3}$$

In plain words, the n-multigram model explores the segmentations compatible with the given string, restricting the length of the segments to $n$ at most; it computes the likelihood of each segmentation as if the segments were independent and, finally, approaches the likelihood of the string as the likelihood of the most likely segmentation.

In this work we turned to the n-multigram model to calculate the joint probability of a clinical entity given in a spontaneous non-standard (**spo**) way and a standard (**st**) candidate counterpart (within the ICD).

Given a possible segmentation within the input and output vocabularies, $\Sigma_{\text{spo}} = \{\alpha_i\}_{i=1}^{\text{SPO}}$ and $\Sigma_{\text{st}} = \{\beta_i\}_{i=1}^{\text{ST}}$ respectively, $\mathbf{s} = s_1^q$ can be seen as a bilingual segmentation compatible with both strings. That is, each element in the segmentation comprises two sub-strings (spo, st) as expressed in (4) where $j_i$ and $k_i$ are the segmentation cut-points in the spontaneous non-standard and standard strings respectively, and particularly, $j_0 = 0 = k_0$.

$$\mathbf{s} = s_1^q = \left(\alpha_{1+j_{i-1}}^{j_i}, \beta_{1+k_{i-1}}^{k_i}\right)_{i=1}^{q} \tag{4}$$

Note that each sub-string length within a given segment $s_i$ is not necessarily the same ($j_i - j_{i-1}$ is not necessarily equal to $k_i - k_{i-1}$) capturing the fact that the length of a spontaneous string usually differs from its standard counterpart (as in examples itemized in page 3).

Weighted Finite-State Transducers (WFST) [12,11,13,14] represent easily a co-segmentation framework [15]. The criterion is to find a path through the transducer compatible with the spontaneous string that minimizes the cost of the path finding the most likely standard string ($\widehat{\mathbf{w}}_{\text{st}}$) given the spontaneous string ($\mathbf{w}_{spo}$). We made use of Phonetisaurus [16], an open source WFST-based Grapheme-to-Phoneme conversion toolkit that provides efficient algorithms such as EM sequence alignment and several decoding techniques.

Note that the length of the non-standard and standard strings ($\mathbf{w}_{\text{spo}} = \alpha_1^{j_1}...\alpha_{1+j_{q-1}}^{j_q}$ and $\mathbf{w}_{\text{st}} = \beta_1^{k_1}...\beta_{1+k_{q-1}}^{k_q}$ respectively) being encompassed by the segmentation might differ (the value of the last index in the non-standard string, $j_q$, might differ from $k_q$, the value of the last index in the standard string).

$$\hat{\mathbf{w}}_{\text{out}} = \arg\max_{\mathbf{w}_{\text{out}}} \ell(\mathbf{w}_{\text{inp}}, \mathbf{w}_{\text{out}}) \tag{5}$$

With the likelihood of the sequence of tokens approached as shown in (6) (which is the generalization from multigrams, what we showed in (2), to joint multigrams) where, for simplicity, $\mathscr{S}$ refers to $\mathscr{S}_{(\mathbf{w}_{\text{inp}},\mathbf{w}_{\text{out}},n)}$, and given the segmentation $\mathbf{s} = (j_1, k_1), \ldots, (j_q, k_q)$ we rewrote the likelihood in expression in (5) in a simpler manner as (7).