Impact of real data from electronic health records on the classification of diagnostic terms

Alicia Pérez
IXA Taldea
(UPV-EHU)

Koldo Gojenola
IXA Taldea
(UPV-EHU)

Maite Oronoz
IXA Taldea
(UPV-EHU)

Arantza Casillas
IXA Taldea
(UPV-EHU)

Abstract

This work tackles Electronic Health Record (EHR) classification according to their Diagnostic Terms (DTs) following the standard International Classification of Diseases-Clinical Modification (ICD-9-CM). To do so, we explore text mining relying on a wide variety of data from both standard catalogues, such as the ICD-9-CM and SNOMED-CT; and, what it was proven even more effective, real data sources, such as EHRs.

The models we put forward to deal with this problem are Finite-State Transducers (FSTs). The aim behind FSTs would be not only to accept exact terms in the ICD-9-CM but also alternative variants. To be precise, a series of FSTs were defined to carry out a soft-matching process between DTs written in natural language and those in the standard form as in the ICD-9-CM catalogue.

1 Introduction

The Clinical Documentation Service of the Galdakao-Usansolo Hospital (a hospital attached to the Spanish Ministry of Health, Social Services and Equality) is interested on automatising the classification of Electronic Health Records (EHRs). EHRs include several fields such as: a description of the patient's details, antecedents, procedures and methods of administration of medicines, and Diagnostic Terms (DTs). It is the DTs that serve as the classification key to classify EHRs according to the World Health Organisation's 9th Revision of the International Classification of Diseases - Clinical Modification (ICD-9-CM)[1]. The goal of this work is to develop a system to automatically classify DTs in an attempt to alleviate the work load by the Clinical Documentation Service but never at the expense of precision. This task presents the following challenges:

1. Natural language in EHRs vs. medical jargon in ICD-9-CM
2. Large-scale classification problem: including more than 14×10^3 different classes
3. Working towards a 100% precision

1.1 Related work

A large number of sophisticated machine learning algorithms have been applied to the task of DT classification. Ferrao et al. (2012) used a commercial system based on either Naive-Bayes or decision trees to tackle multi-label classification of EHRs restricted to the Internal Medicine department.

The top systems in the 2007 Computational Medicine Challenge have benefited from incorporating domain knowledge of free-text clinical notes, such as negation, synonymy and hyperonymy, either as hand-crafted rules in a symbolic approach, or as carefully engineered features in a machine learning component: (Goldstein et al., 2007; Crammer et al., 2007; Aronson et al., 2007; Patrick et al., 2007). Yet, this shared task involved the assignment of ICD-codes to radiology reports written in English from a reduced set of 45 codes (Pestian et al., 2007). By contrast, we focus on the entire scope of the ICD-9-CM catalogue.

Most of the systems described in the literature were developed for English. Looking at other languages, Metais et al. (2007) reported a system to classify medical reports in French.

[1] The reader might be aware of the fact that for English other codification systems (such as ICD-10) are also reported in the literature, nevertheless, it is the ICD-9-CM the one being currently used by the Spanish Health System even though it is foreseen to move to ICD-10 in the near future.

2 Methods: Finite-State Transducers

Finite-State Automata (FSA) serve to the purpose of recognising regular grammars (Chomsky, 1959). A grammar is used to either generate or parse the strings accepted in the language recognised by the FSA. In our medical domain the DTs in the ICD-9-CM catalogue represent the set of acceptable strings within a formal language with a particular syntax. Thus, inferring the grammar underlying the DT domain would help to assess whether a given string could be considered or not appropriately expressed in that language.

Finite-State Transducers (FSTs) are an extension of FSAs that encompasses two languages: input and output. FSTs serve to analyse an input string and associate an output string (in case that the input is acceptable in the source language). That is, FSTs serve to map from one language to the other. The nature of the FSTs does not allow to accept any string out of the language, and this property strives towards a high precision.

2.1 Implementation

In brief, the system is designed as a composition of three FSTs: lexicon, normalisation and generation. The FSTs were next integrated on a priority union basis. This operation allows a wide search while it tries to stick as possible to the input. Besides, it rejects some strings, meaning that it reveals ill-formed DTs. All the FSTs as well as their operations were implemented through Foma (Hulden, 2009). Foma is a freely available toolkit that allows to build finite-state transducers and also includes efficient parsing functions. Besides, it supports imports from, and exports to, other toolkits, such as Xerox's XFST (Beesley and Karttunen, 2003), AT&T (Mohri et al., 2003) and OpenFST (Riley et al., 2009). Next we provide some details of each FST:

1. **FST-Lexicon:** it compiles the reference collection of allowed (DT, ICD-code) pairs, that is, the lexicon of the application. This FST is automatically built by Foma from the set of pairs allowed. The data-sets involved in the lexical model came from two sources:
   - ICD-9-CM: consists of more than $14,435$ different (DT, ICD-code) pairs not restricted to a single clinical domain.
   - EHRs in Spanish: a set of more than 28,000 (DT, ICD-code) pairs with DTs written by doctors and coded by experts in EHRs that allows supervised classification.
2. **FST-Normalisation:** it carries out elementary pre-processing operations. The goal is to get all the inputs re-cased, to get rid of written accents and other punctuation marks that are considered as noisy. This FST was built from rules and compiled as an FST by Foma. An example of the rules underlying this FST is given in Figure 1a.
3. **FST-Generation:** it allows to generalise the reference lexicon by means of synonyms, acronyms, etc. As a result, it allows to generate new alternatives for the DTs. This FSTs implements rules to check punctuation marks, to allow number variation (to create singular and plural forms for a given DT in the reference), the omission and equivalence of some prepositions, either expand abbreviations, synonyms of the reference according to SNOMED-CT, optional replacement in a given context, composition, union, projection,etc. For exemplification purposes, some of these rules are shown in a very simplified manner in Figure 1b.

Let us show in an example the procedure by which the system makes it possible the automatic assignment of the correct ICD-code, 185, to the DT "Ca. prostata" used in an EHR. In the ICD-9-CM the term encoded with 185 is "Neoplasia maligna de la próstata". Hence, an exact lookup operation would have been unproductive. Nevertheless, the soft-matching operations implemented through the proposed FST are able to find the required term, and accordingly, provide the corresponding ICD-code. As a first step, both terms (the DT and the one in the ICD list) are normalised by the FST-Normalisation that was defined from the set of rules denoted as `Accents` and `Low2Upp` (see Figure 1a). The normalisation step yields "CA. PROSTATA" and "NEOPLASIA MALIGNA DE LA PROSTATA". After that, the FST-Generation proceeds with the generation of several alternatives: the `AltCa` rule enables the equivalence of several alternatives, such as "CA." and "NEOPLASIA MALIGNA". Hence, this enables to parse "CA. PROSTATA" as "NEOPLASIA MALIGNA PROSTATA". Finally, the `Preps` rule adds the prepositions, leading to the standard

```
Accents [á -> a].o.[é -> e].o.  ...  .o.  [ú -> u];
Low2Upp [a -> A].o.[b -> B] .o.  ...  .o.  [z -> Z];
⋮       ⋮
```

(a) Normalisation

```
Pl_I    [S|ES] (->) "" || Upper _ [.#.  | "." | ","];
Pl      [..]  (->) ([S|ES]) || Upper _ [.#.  | "." ];
R4      IV (->) "4" || " " _ [.#.|"."|" "];
Preps   [..]  (->) [de |del| de la |con |por ]||" "_;
AltCa   [NEOPLASIA MALIGNA|CA.|ADENOCARCINOMA|...];
EquivCa [AltCa:AltCa];
⋮       ⋮
```

(b) Generation

Figure 1: Rules underlying the FSTs involved: FST-Normalisation and FST-generation

term in the ICD list "NEOPLASIA MALIGNA DE LA PROSTATA" from the DT in the EHR "CA. PROSTATA".

The FSTs were arranged with a priority union in such a way that each FST contributed with additional capabilities to the previous one. The transducers were composed in such a way that the most simple transducer was looked-up first and the one allowing the higher variability last. That is, a priority union is applied to compose the different transducers.

3 Experimental results

For this task it is preferred to get accurate results with high precision even at the expense of low coverage. Hence, the system allows rejections whenever the input DT does not match any of the alternatives allowed in the language accepted by the FST. That is, all the instances that did not soft-match a DT in the FST are left unclassified and this is why we are not referring to our system as a fully automatic classification system but as a classification support system, instead.

Accordingly, for a given DT there are three possible outcomes:

Reject: the DT was not assigned any code by the system because the input DT did not soft-match any of the accepted alternatives in the FST. That is, there was not any path in the transducer accepting the source string.

Miss: the DT was assigned a code by the system that did not match the manually assigned ICD-code.

Hit: the DT was assigned a code that matched the one in the reference.

The performance of the FST, shown in Table 1, was assessed using a 5-fold cross validation on the EHR set of 28,000 (DT, ICD-code) pairs, while including also the ICD-9-CM set to feed the FST-Lexicon.

In order to make clear the relevance of both the nature of the seed lexicon and the generation operation, we made a baseline experiment: the lexicon consisted only of the standard ICD-9-CM set of pairs and while normalisation operation was allowed, we did not allow for any generation. Through this *baseline* we meant to measure the number of DTs written by doctors nearly as in the standard ICD-9-CM. Although the ICD-9-CM is composed of 14,435 different pairs, the number of hits achieved was 7.1%. Moreover, allowing next the generation operation on the same lexicon, the hits represent the 8.1%, the rejections the 89.0% and the misses the 2.9%. Comparing this baseline with the results in Table 1, the conclusion drawn is that the aid of real EHRs seems to be of much benefit in what comes to feeding the lexicon of the FST.

Evaluation	**Rejections**	**Misses**	**Hits**
automatic	12.0%	1.2%	86.8%

Table 1: Performance of the FST.

3.1 Impact of real data on performance

Having incorporated EHRs to the allowed lexicon provided excellent results with respect to the baseline. Hence, it seemed of interest to quantitatively assess the impact of including more and more instances from EHRs, which is, precisely, one of the hubs of this paper.

The aim is to learn a regression model that would predict the effect of adding further data on the coverage. To do so, more and more instances from EHRs were progressively added to the lexicon and the improvements in terms of coverage were evaluated. A polynomial regression on the evaluation data was carried out showing the following approximated relation:

$$y \approx f(x) = a_2 \cdot x^2 + a_1 \cdot x + a_0 \qquad (1)$$

being:

x the size of the (DT, ICD-code) pairs from EHRs used to feed the FST-Lexicon, presented in logarithmic scale.

y the number of rejections provided by the FST, expressed as a percentage.

to be precise:

$$x = ln(|\mathcal{C}|) \quad (2)$$

$$y = \frac{|\mathcal{R}| * 100}{|\mathcal{R}|} \quad (3)$$

On this basis, a quadratic polynomial predictive model presented in eq. (1) was derived with the following coefficients:

$$a_2 = 1.57 \qquad a_1 = -37.5 \qquad a_0 = 226 \quad (4)$$

These results, represented in Figure 2, show that even a small corpus would represent a leverage to gain on coverage for similar tasks.

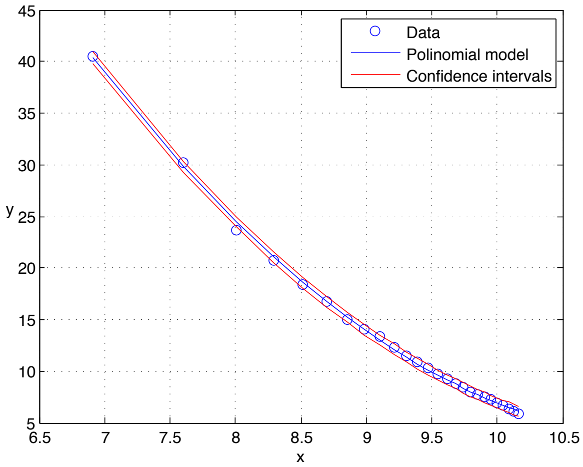

Figure 2: The number of rejections as a percentage (in the ordinate) with respect to the size of the corpus in logarithmic scale (abscissa). Experimental results are represented as circles. The quadratic polynomial function proposed in eq. (1) is represented together with its confidence interval by the curve and its upper and lower bounds.

The experimental results show that the corpus plays a core role on the performance of the system. While the standard ICD list showed to be of help, significantly better results were obtained extracting the lexicon from previously classified DTs written in EHRs. The impact of adding more and more DTs from previous EHRs to the corpus has shown to reduce the number of unclassified DTs in a logarithmic basis. Moreover, as a side effect the precision was also improved.

4 Concluding remarks and future work

In this work we present a system to classify diagnostic terms in Spanish according to the ICD-9-CM standard. The approach was based on the representation of a corpus of (DT, ICD-code) pairs in terms of FSTs that would parse an input DT into an output ICD-code.

The experimental results showed that the corpus played a core role on the performance of the system. The role played by the corpus opens another line of research: possibly lower amounts of data could be used with similar performance making use of adaptive models for different user-profiles (writing styles, use of abbreviations, etc.).

To sum up, the contribution of this paper are:

1. Large-scale and high precision automatic DT classification: the main contribution of this work is a high precision automatic classification of DTs in EHRs according to the ICD-9-CM reference. We propose the use of the FST framework, that allows not only to do an exact lookup but also a soft-matching within the lexicon or a set of positive samples.

2. Quantification of the benefits of real data: we propose the use of previously classified corpus in order to enhance the matching process adding DTs written differently to the standard.

3. Development of medical resources in Spanish: to the authors' knowledge this is the first attempt using all the codes in the ICD list in Spanish and rule-based pattern recognition approach. In addition, we contributed with an underlying process of acquisition and also with a pre-processing of valuable lexical resources within the medical domain in Spanish.

Future work will focus on those DTs that were rejected by the system (and thus, left unclassified) in an attempt to gain coverage. Together with FSTs, other strategies, such as support vector machines shall be explored. While this work was presented as an automatic classification approach, since the goal is to arise a 100% precision, it seems of interest to explore the unclassified DTs through interactive pattern recognition approaches (Toselli et al., 2011). This is can also be achieved through FSTs, since they were proven efficient in computer-aided tasks.

Acknowledgments

We would like to thank the Hospital Galdakao-Usansolo for their contributions and support, in particular to Javier Yetano, responsible of the Clinical Documentation Service.

This work was supported by the Department of Industry of the Basque Government (IT344-10, S-PE12UN114), the University of the Basque Country (GIU09/19), the Spanish Ministry of Science and Innovation (TIN2010-20218, TIN2012-38584-C06-02).

References

[Aronson et al.2007] A.R. Aronson, O. Bodenreider, D. Demner-Fushman, K.W. Fung, V.K. Lee, J.G. Mork, A. Neveol, L. Peters, and W.J. Rogers. 2007. From indexing the biomedical literature to coding clinical text: Experience with MTI and machine learning approaches. In *Proceedings of the Workshop on BioNLP*, pages 105–112.

[Beesley and Karttunen2003] Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications,.

[Benesch et al.1997] C Benesch, DM Witter, AL Wilder, PW Duncan, GP Samsa, and DB Matchar. 1997. Inaccuracy of the international classification of diseases (icd-9-cm) in identifying the diagnosis of ischemic cerebrovascular disease. *Neurology*, 49(3):660–664.

[Chomsky1959] Noam Chomsky. 1959. On certain properties of formal grammars. *Information and Control*, 2(2):137—167.

[Crammer et al.2007] K. Crammer, M. Dredze, K. Ganchev, P.P. Talukdar, and S. Carroll. 2007. Automatic code assignment to medical text. In *Proceedings of the Workshop on BioNLP*, pages 129–136.

[Ferrao et al.2012] J.C. Ferrao, M.D. Oliveira, F. Janela, and H.M.G. Martins. 2012. Clinical coding support based on structured data stored in electronic health records. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on*, pages 790–797.

[Goldstein et al.2007] I. Goldstein, A. Arzumtsyan, and O. Uzuner. 2007. Three approaches to automatic assignment of ICD-9-CM codes to radiology reports. In *Proceedings of the AMIA Annual Symposium*, pages 279–283.

[Hulden2009] Mans Hulden. 2009. Foma: a Finite-State Compiler and Library. In *European Association for Computational Linguistics*, pages 29–32. The Association for Computational Linguistics (ACL).

[Lita et al.2008] Lucian Vlad Lita, Shipeng Yu, Radu Stefan Niculescu, and Jinbo Bi. 2008. Large scale diagnostic code classification for medical patient records. In *Third International Joint Conference on Natural Language (IJCNLP)*, pages 877–882, Hyderabad, India, January. The Association for Computer Linguistics.

[Metais et al.2007] Elisabeth Metais, Didier Nakache, and Jean-François Timsit. 2007. Automatic classification of medical reports, the cirea project. In *Proceedings of the 5th WSEAS International Conference on Telecommunications and Informatics*, pages 354–359.

[Mohri et al.2003] Mehryar Mohri, Fernando C. N. Pereira, and Michael D. Riley. 2003. AT&T FSM LibraryTM – Finite-State Machine Library.

[Patrick et al.2007] J. Patrick, Y. Zhang, and Y.Wang. 2007. Evaluating feature types for encoding clinical notes. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 218–225.

[Pestian et al.2007] John P. Pestian, Chris Brew, Pawel Matykiewicz, D. J Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch. 2007. A shared task involving multi-label classification of clinical free text. In *Biological, translational, and clinical language processing*, pages 97–104, Prague, Czech Republic, June. Association for Computational Linguistics.

[Riley et al.2009] Michael Riley, Cyril Allauzen, and Martin Jansche. 2009. OpenFST: An open-source, weighted finite-state transducer library and its applications to speech and language. In *Proceedings of Human Language Technologies. Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 9–10. Association for Computational Linguistics.

[Toselli et al.2011] Alejandro H. Toselli, Enrique Vidal, and Francisco Casacuberta. 2011. *Multimodal Interactive Pattern Recognition and Applications*. Springer.