First approaches on Spanish medical record classification using Diagnostic Term to class transduction

A. Casillas$^{(1)}$, **A. Díaz de Ilarraza**$^{(2)}$, **K. Gojenola**$^{(2)}$, **M. Oronoz**$^{(2)}$, **A. Pérez**$^{(2)}$
$^{(1)}$ Dep. Electricity and Electronics
$^{(2)}$ Dep. Computer Languages and Systems
University of the Basque Country (UPV/EHU)
arantza.casillas@ehu.es

Abstract

This paper presents an application of finite-state transducers to the domain of medicine. The objective is to assign disease codes to each Diagnostic Term in the medical records generated by the Basque Health Hospital System. As a starting point, a set of manually coded medical records were collected in order to code new medical records on the basis of this set of positive samples. Since the texts are written in natural language by doctors, the same Diagnostic Term might show alternative forms. Hence, trying to code a new medical record by exact matching the samples in the set is not always feasible due to sparsity of data. In an attempt to increase the coverage of the data, our work centered on applying a set of finite-state transducers that helped the matching process between the positive samples and a set of new entries. That is, these transducers allowed not only exact matching but also approximate matching. While there are related works in languages such as English, this work presents the first results on automatic assignment of disease codes to medical records written in Spanish.

1 Introduction

During the last years an exponential increase in the number of electronic documents in the medical domain has occurred. The automatic processing of these documents allows to retrieve information, helping the health professionals in their work. There are different sort of valuable data that help to exploit medical information. Our framework lays on the classification of Medical Records (MRs) according to a standard. In our context, the MRs produced in a hospital have to be classified with respect to the World Health Organization's 9th Revision of the International Classification of Diseases[1] (ICD-9). ICD-9 is designed for the classification of morbidity and mortality information and for the indexing of hospital records by disease and procedure. The already classified MRs are stored in a database that serves for further classification purposes. Each MR consists of two pieces of information:

Diagnostic Terms (DTs): one or more terms that describe the diseases corresponding to the MR.

Body-text: a description of the patient's details, antecedents, symptoms, adverse effects, methods of administration of medicines etc.

Even though the DTs are within a limited domain, their description is not subject to a standard. Doctors express the DTs in natural language with their own style and different degrees of precision. Usually, a given concept might be expressed by alternative DTs with variations due to modifiers, abbreviations, acronyms, dates, names, misspellings or style. This is a typical problem that arises in natural language processing due to the fact that doctors focus on the patients and not so much on the writing of the MR. On account of this, there is ample variability in the presentation of the DTs. Consequently, it is not a straightforward task to get the corresponding ICD-codes. That is, the task is by far more complex than a standard dictionary lookup.

[1] http://www.cdc.gov/nchs/icd/icd9.htm

The Basque Health Hospital System is concerned with the automatization of this ICD-code assignment task. So far, the hospital processes the daily produced documents in the following sequence:

1. **Automatic:** exact match of the DTs in a set of manually coded samples.

2. **Semi-automatic:** through semantic match, ranking the DTs by means of machine-learning techniques. This stage requires that experts select amongst the ranked choices.

3. **Manual:** the documents that were not matched in the previous two stages are examined by professional coders assigning the codes manually.

The goal of this paper is to bypass the variability associated to natural language descriptions in an attempt to maximize the proportion of automatically assigned codes, as the Hospital System aims to expand the use of the automatic codification of MRs to more hospitals. According to experts, even an increase of 1% in exact match would represent a significant improvement allowing to gain time and resources.

Related work can be found in the literature. For instance, Pestian et al. (2007) reported on a shared task involving the assignment of ICD-codes to radiology reports written in English from a reduced set of 45 codes. In general it implied the examination of the full MR (including body-text). In our case, the number of ICD-codes is above 1,000, although we restrict ourselves to exact and approximate match over the diagnoses.

Farkas and Szarvas (2008) used machine learning for the automatic assignment of ICD-9 codes. Their results showed that hand-crafted systems could be reproduced by replacing several laborious steps in their construction with machine learning models.

Tsuruoka et al. (2008) presented a system that tried to normalize different variants of the terms contained in a medical dictionary, automatically getting normalizing rules for genes, proteins, chemicals and diseases in English.

The contribution of this work is: i) to collect manually coded MRs in Spanish; ii) to approximate transduction with finite-state (FS) models for automatic MR coding and, iii) to assess the performance of the proposed FS transduction approaches.

2 Approximate transduction

As it was previously mentioned, there are variations regarding the DT descriptions due to style, missspells, etc. Table 1 shows several pairs of DT and ICD-codes within the collected samples that illustrate some of those variations.

	DT	**ICD**
1	*Adenocarcinoma de prostata*	*185*
2	*Adenocarcinomas próstata.*	*185*
3	*Ca. prostata*	*185*
4	*CÁNCER DE PROSTATA*	*185*
5	*adenocarcinoma de pulmon estadio IV*	*1629*
6	*CA pulmón estadio 4*	*1629*
7	*ADENOCARCINOMA PANCREAS*	*1579*

Table 1: Examples of DTs and their ICD-codes.

There are differences in the use of uppercase/lower case; omissions of accents; use of both standard and non-standard abbreviations (e.g. *ca.* for both *cáncer* and *adenocarcinoma*); punctuation marks (incidental use of full-stop as commas, etc.); omission of prepositions (see rows 1 and 2); equivalence between Roman and Arabic numerals (rows 5 and 6). Due to these variations, our problem can be defined as an approximate lookup in a dictionary.

2.1 Finite-state models

Foma toolkit was used to build the FS machines and code the evaluation sets. Foma (Hulden, 2009) is a freely available[2] toolkit that allows to both build and parse FS automata and transducers. Foma offers a versatile layout that supports imports/exports from/to other tools such as: Xerox XFST (Beesley and Karttunen, 2003), AT&T (Mehryar Mohri and Riley, 2003), OpenFST (Riley et al., 2009). There are, as well, outstanding alternatives such as HFST (Lindén et al., 2010). Refer to (Yli-Jyrä et al., 2006) for a thorough inventory on FS resources.

The FS models in Figure 1 perform the conversions necessary to carry out a soft match between the dictionary entries and their variants.

- First, we define the transducer `Accents` that takes into account the correspondences between standard letters and their versions using accent text marks.

[2] `http://code.google.com/p/foma`

```
define  Accents      [a:á|e:é|i:í|o:ó|u:ú|...];
define  Case         [a:A|b:B|c:C|d:D|e:E|f:F|...];
define  Spaces       [..]  (->) " " || [.#.  | "."] _ ,_ .#.;
define  Punctuation  ["."|"-"|" "]:["."|"-"|" "];
define  Plurals      [..]  -> ([s|es]) || _ [.#.  | "." | " "];
define  PluralsI     [s|es] (->) "" || _ [.#.  | "." | ","| " "];
define  Preps        [..]  (->) [de |del |con |por ] || " "_;
define  Disease      [enf|enf.|enfermedad]:[enf|enf.|enfermedad];
define  AltCa        [tumor|ca|ca.|carcinoma|adenocarcinoma|cáncer];
define  TagNormCa    AltCa:AltCa;
define  AltIzq       [izquierdo|izquierda|izq|izq.|izqda|izqda.|
                     izqdo|izqdo.|izda|izda.|izdo|izdo.];
define  TagNormIzq   AltIzq:AltIzq;
```

Figure 1: A simplified version in Foma source code of the regular expressions and transducers used to bypass several sources of distortion within the DTs in order to parse variations of unseen input DTs.

- The expression Case matches uppercase and lowercase versions of the DTs.
- There is a set of transducers (Spaces, Punctuation, Plurals and PluralsI) that deal with the addition or deletion of spaces and separators (as full-stop, comma, and hyphen) between words or at the end of the DT.
- Prepositions. Many DTs can be differentiated by the use or absence of prepositions, although they correspond to the same ICD-code. For that reason, we designed a transducer that inserts or deletes the prepositions from a reduced set that were identified by inspection of the training set. In this way, expressions as "*Adenocarcinoma prostata*" and "*Adenocarcinoma de prostata*" can be mapped to each other.
- Tag Normalization of synonyms, variants and abbreviations. The examination of the DTs in the training set revealed that there were several terms used indistinctly, including synonyms and different kinds of variants (masculine and feminine) and abbreviations. For example, the words *adenocarcinoma, adenoca., carcinoma, ca, ca. and cancer* serve to name the same disease. There are also multiple variants of left/right, indicating the location of an illness, that do not affect the assignment of the ICD-code (e.g. *izquierdo, izq., izda.*).

Finally, all the FS transducers were composed into a single machine that served to overcome all the sources of distortion together.

3 Experimental results

To begin with, coded MRs produced in the hospital throughout 12 months were collected summing up a total of 8,020 MRs as described in Table 2. Note that there are ambiguities in our data-set since there are 3,313 different DTs that have resulted in 3,407 (DT, ICD-code) different pairs (as shown in Table 2). That is, the same DT was not always assigned the same ICD-code.

	DT	**ICD-code**
entries	8,020	
different entries	3,407	
different forms	3,313	1,011

Table 2: The data-set of (DT, ICD-code) pairs.

Next, the data-set was shuffled and divided into 3 disjoint sets for training, development and test purposes as shown in Table 3.

	train	**dev**	**test**
entries	6,020	1,000	1,000
different entries	2,825	734	728

Table 3: The data-set shuffled and divided into 3 sets

Using the set of mappings derived from the training set we performed the experiments on the development set. After several rounds of tuning the system, the resulting system was applied to the test set.

		PERCENTAGE OF UNCLASSIFIED DTs					
TRAIN	**EVAL-SET**	**exact-match**	**+ case-ins.**	**+ punct.**	**+ plurals**	**+preps.**	**+ tag-norm.**
train	**dev**	30.6	27.0	25.2	24.4	23.9	23.2
train	**test**	29.8	26.7	25.1	24.8	24.3	23.2
train+dev	**test**	27.7	24.5	23.0	22.9	22.5	21.4

Table 4: Performance of different FS machines in terms of the percentage of unclassified entries. All the classified entries were correctly classified, yielding, as a result, a precision of 100%.

Given a DT, the goal is to find its corresponding ICD-code despite the variations. Different FS approaches (described in Section 2.1) were proposed to bypass particular sources of noise in the DT. Their performance was assessed by means of the percentage of unclassified DTs, as summarized in Table 4. Note that the lower the number of unclassified DTs the better the performance. In each of the three rows of Table 4 the results of different experimental setups are shown: in the first two rows the training set was used to build the models and either the development or the test set was evaluated in their turn; in the third row, both the training and the development sets were used to build the model and the test set was evaluated. The impact of adding progressively the FS machines built to tackle particular sources of noise is shown by columns. Thus, the results of the last column represent the performance of the transducer allowing exact-match search together with case-insensitive search, bypassing punctuation marks, allowing plurals, bypassing prepositions and allowing tag-normalization. The composition of each transducer outperforms the previous result, yielding an improvement on the test of 6 absolute points over the exact-match baseline, from 27.7% to 21.4%. As it can be derived from the first column of Table 4 the test set contributed to the training+development set with %27.7 of new DTs.

Overall, the FSMs progressively improved the results for the three series of experiments carried out in more than 6%. As a result, less and less DTs are left unclassified. In other words, the FS machines tackling different sources of errors contribute to assign ICD-codes to previously unassigned DTs.

A manual inspection over the results associated to the evaluation of the development set (focus on the first row of Table 4) showed that all the DTs were correctly classified according to the training data. Overall, the resulting transducer was unable to classify 232 DTs out of 1,000 (see last column in first row). Among the unclassified DTs, 10 out of 232 were due to misspellings: e.g. *cic atriz* (instead of *cicatriz*), *desprendimineot* (instead of *desprendimiento*). In fact, spelling correction reported improvements in related tasks (Patrick et al., 2010). The remaining DTs showed wider variations in their forms, as unexpected degree of specificity (e.g. named entities), spurious dates or numbers.

4 Conclusions

Medical records in Spanish were collected yielding a data set of 8,020 DT and ICD-code pairs. While there are a number of references dealing with English medical records, there are few for Spanish.

The goal of this work was to build a system that given a DT it would find its corresponding ICD-code as in a standard key-value dictionary. Yet, the DTs are far from being standard since they contain a number of variations. We proposed the use of several FS models to bypass different variants and allow to provide ICD-codes even when the exact DT was not found. Each source of variations was tackled with a specific transducer based on handwritten rules. The composition of each machine improved the performance of the system gradually, leading to an improvement up to 6% in accuracy, from 27.7% unclassified DTs with the exact-match baseline to 21.4% with the tag-normalization transducer.

Future work will focus on the unclassified DTs. Together with FS models, other strategies shall be explored. Machine-learning strategies in the field of information retrieval might help to make the most of the piece of information that was here discarded (i.e. the body-text). All in all, regardless of the approach, the command in this MR classification context is to get an accuracy of 100%, possibly through the interactive inference framework (Toselli et al., 2011).

Acknowledgments

Authors 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 research was supported by the Department of Industry of the Basque Government (IT344-10, S-PE11UN114, GIC10/158 IT375-10), the University of the Basque Country (GIU09/19) and the Spanish Ministry of Science and Innovation (MICINN, TIN2010- 20218).

References

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

[Farkas and Szarvas2008] Richárd Farkas and György Szarvas. 2008. Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics.*, 9 (Suppl 3): S10.

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

[Lindén et al.2010] Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. 2010. HFST tools for morphology – an efficient open-source package for construction of morphological analyzers.

[Mehryar Mohri and Riley2003] Fernando C. N. Pereira Mehryar Mohri and Michael D. Riley. 2003. AT&T FSM LibraryTM – Finite-State Machine Library. `www.research.att.com/sw/tools/fsm`.

[Patrick et al.2010] Jon Patrick, Mojtaba Sabbagh, Suvir Jain, and Haifeng Zheng. 2010. Spelling correction in clinical notes with emphasis on first suggestion accuracy. In *2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2010) LREC*. ELRA.

[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: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, pages 9–10, Boulder, Colorado, May. Association for Computational Linguistics.

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

[Tsuruoka et al.2008] Yoshimasa Tsuruoka, John McNaught, and Sophia Ananiadou. 2008. Normalizing biomedical terms by minimizing ambiguity and variability. *BMC Bioinformatics*, 9(Suppl 3):S2.

[Yli-Jyrä et al.2006] A. Yli-Jyrä, K. Koskenniemi, and K.. Lindén. 2006. Common infrastructure for finite-state based methods and linguistic descriptions. In *Proceedings of International Workshop Towards a Research Infrastructure for Language Resources.*, Genoa, May.