From Episodes of Care to Diagnosis Codes: Automatic Text Categorization for Medico-Economic Encoding

Patrick Ruch[ab], Julien Gobeill[a], Imad Tbahriti[a], Antoine Geissbühler[a]

[a]*Medical Informatics Service, University and University Hospitals of Geneva, Geneva*
[b]*College of Library Sciences, University of Applied Sciences Western Switzerland, Geneva, Switzerland*
Contact : patrick.ruch@sim.hcuge.ch

Abstract

We report on the design and evaluation of an original system to help assignment ICD (International Classification of Disease) codes to clinical narratives. The task is defined as a multi-class multi-document classification task. We combine a set of machine learning and data-poor methods to generate a single automatic text categorizer, which returns a ranked list of ICD codes. The combined ranking system currently obtains a precision of 75% at high ranks and a recall of about 63% for the top twenty returned codes for a theoretical upper bound of about 79% (inter-coder agreement). The performance of the data-poor classifier is weak, whereas the use of temporal features such as anamnesis and prescription contents results in a statistically significant improvement.

INTRODUCTION

We present an attempt to develop a medical encoding system to help a staff of about 20 professional coders in the assignment of ICD (International Classification of Diseases) codes. The study focuses on hospitalized episodes, whose billing if fully based on DRGs (Diagnosis Related Groups) at the University Hospitals of Geneva, a large healthcare institutions of more than 2000 beds. Medico-economics encoding in the institution is performed at the level of episodes of care. An episode of care can gathers several documents (e.g. admission notes, follow up notes, laboratory and test results, radiology reports,), but is normally concluded by a discharge summary or a surgery report. Thus, the medical encoding task we are studying can be sequentially described as 1) a document selection task, 2) a paragraph selection task, 3) a manual search in the ICD book assisted by an advanced browsing tool in the terminology as well as in previously coded cases. As legally defined in the institution, the encoding task is a multi-document, multi-class automatic text categorization task [23].

Automatic Text Categorization (ATC) is formally defined as a classification task: given a textual input, the categorizer should return a list of categories, which are supposed to provide non-ambiguous machine-readable information about the input text. Typical applications are for example: document routing (e.g. Reuters and Bloomberg financial newswires [1]), document indexing by librarians (e.g. Medical Subject Headings assigned to articles in the MEDLINE digital library [2]). Application of ATC to the automatic assignment of medical descriptors has been largely studied in the context of document indexing [3], to help cross-language information retrieval on the medical web [4] or for functional prediction in proteomics [5]. Applications of these methods to encoding of patient-related data are more seldom: thus, [6] attempts to assign SNOMED categories to clinical contents, while Pakhomov and al. 2006 [7] categorize clinical narratives into ICD-9 categories.

From a methodological perspective, it is possible to separate computer-based text categorization technologies in two subsets:

1. *retrieval based on string matching*, which assign descriptors to texts based on some shared lexical features;

2. *empirical learning of a classification model* from a training set of texts with their associated concepts.

In the former approach, the descriptors are indexed and each indexing unit (e.g. a word, a stem or a phrase) receives a specific weight, while for the latter, a more complex model of the data is built up in order to provide text-concept associations beyond strict features sharing. Word-based matching approaches, which include vector-space [24] and pattern matching engines [25], are often presented as weak categorization methods, see e.g. [26][27][28], because associations between text and categories are based on simple string matching strategies, but in several practical situations learning approaches cannot be applied. In [5], the authors argue that the use of data-poor ranking-based methods could be of interest in the biomedical domain because these methods are computationally cheaper and because they are able to perform categorization tasks with extremely little training data, as it is frequent in life sciences due to the massive growth of categories in this scientific field. In contrasts with this studies, we hypothesize that for diagnosis encoding, data-intensive approaches applied on a rather static category set like ICD – no more than a handful of cate-

gory is added per year – should overcome data-poor approached. Therefore, we intend to apply and evaluate both approaches to perform automatic diagnosis encoding. In the proposed experiments, the document selection is performed based on a master list of document type, which indicates whether a particular document type in the patient file is necessary, useful, or not useful for medical encoding. The paragraph selection focuses on the three following free-text sections: anamnesis, diagnosis and prescription.

Data and Metrics

Data resources used to build up the system and our experiments are the following: the French ICD-10, a French thesaurus [12, 17], and the coding data from the institution data warehouse [18] collected between the 1st of January 2004 (total: ~733'484 codes; ~10511 unique codes) and the end of the year 2006. From these data, we observe that on average 5.1 codes are assigned per episode of care. From these data, we were able to generate the following distribution of paragraph-code association pairs: {anamnesis, ICD codes} = 18655 champs (23%); {prescription, ICD codes} = 61372 champs (78%) ; {diagnostics, ICD codes} 78756 (100%). It means that the diagnosis field is available for every episode of care, while some of them do not contain any anamnesis or prescription field. In this presentation, all types of diagnosis fields (*anotomo-pathology, co-morbidities, psycho-social*…) are conflated into a generic diagnosis class. The same applies to anamnesis (*anamnesis by systems, interim anamnesis*…) and prescription fields, which are respectively concatenated in two generic classes. Out of our data collection, 800 instances (~1%) are kept for final validation of our system. In our experiments, we do not separate between the main diagnosis and the associated diagnosis.

As usual for classification tasks, our quantitative evaluations are based on precision and recall metrics [19]. Because we design the task as an interactive classification task rather than a fully automatic classification task, we report the precision of the top ranked category predicted by the system together with the recall of the system after twenty categories. As the system is designed to be used to help medical encoders, displaying more than twenty categories does not seem suitable. Although the institution is concerned about both recall (comprehensiveness) and precision (quality), the planned interactive use of the system makes the recall more important than the precision. We indeed expect that a wrongly assigned category could be easily discarded by an expert, whereas missing a relevant category could directly result in an economical cost. In addition to recall at twenty categories (R20), the precision of the top-ranked category (Precision at rank 0 or P0), which measures the effectiveness of the system in a fully-automatic setting, is also provided.

METHODS

We combine several classifiers: a set of supervised learners and a data-poor categorizer. The data poor system covers the full list of valid ICD-10 diagnosis i.e. about 20183, together with their morphological variants (about 10'000 synonyms). The search space of the learning-based system is more limited, since only about 10511 codes are available in the data warehouse. Three independent nearest neighbours were designed for the three different main textual fields available in the clinical narratives: diagnosis, anamnesis, prescription. Finally a combination model is proposed [9]: it combines linearly the output of the three nearest neighbours together with the output of the data-poor classifier, which is only applied to diagnosis fields. More elaborated fusion models, which take differentially into account the local statistical estimate provided by each classifier, in particular the retrieval and categorization status value returned by the ranker are under evaluation.

Data pre-processing

The document collection follows several pre-processing steps in order to obtain a more normalized representation of data: data acquisition, de-identification (or anonymization [22]), format normalization (UTF-16/8, RTF, HTML, PS), quality restoration (misspellings, diacritics), feature selection (stop words, negation handling), and stemming [14]. De-identification is performed mainly to avoid confusing patients and clinicians names with medical diagnosis and procedures (e.g. *Parkinson*, *Donati*…). Since the final application could be potentially made available to other institutions, de-identification is also useful to avoid direct associations between identities and diagnosis.

Data-poor classifier

This classifier use strict lexical, thesaural and linguistically-motivated features, weighted by a simple but robust statistical model. The data-poor framework has been extensively described in [5]. The system used in this study was tuned on a sample of 100 instances. Evaluated on a MEDLINE citation indexing task [29], as well as in density estimation for molecular biology categorization tasks [30], the system obtained highly competitive results in the context of the first BioCreative evaltuation campaign [13]. Unlike the kNN systems, which rely on similarities between a care episode to be coded and a collection of coded episodes, the data-poor classifier is likely to propose codes, which have never been assigned previously. Thus if a tropical *Ebola fever* was to be diagnosed in Geneva, the data-poor system could still propose the related code based on simple lexical similarities.

k-Nearest Neighbours and classifiers fusion

Each kNN, generated from each type of textual data, is tuned by ten fold cross-validation. During the tuning process, but we observed that for defining the neighbours' centroid, distances based on pivoted normalization were particularly effective (formula are given in Table 7). This observation is consistent with those reported for retrieving similar articles in MEDLINE, which showed that pivoted normalization was effective [10, 11, 15].

The fusion model used for these preliminary experiments is inspired by information retrieval [9] and seems particularly adapted to the ranking classification task that we design. It combined linearly the statistical estimates provided by each single classifier. While more advanced combination models exist (e.g. [20, 21]), such ranking-based approaches are particularly robust. They remain effective when training data are sparse and are less likely to be affected by over fitting phenomena.

RESULTS AND DISCUSSION

Not surprising, results reported on the validation set in Tables 2, 3, 4 show that the diagnosis field is the best predictor for ICD codes, with a precision of 74% (P0) and a recall of 59% (R20). The second best is the anamnesis (P0=57, R20=39), while the prescription field achieves a precision of only 49% and a recall of 29%. In Table 5, we observe that our data-poor categorizer does not perform well (P=22% and R20=7%). The combination of these basic classifiers results in a modest improvement regarding precision (from 74% to 75%) and a more significant improvement regarding recall from 59% to 63%, which means a gain of +6.8% ($p < 0.01$).

Mesures	Results
P0	0.74
R20	0.59

Table 2. Results of KNN on Diagnosis data.

Mesures	Results
P0	0.57
R20	0.39

Table 3. Results of KNN on Anamnesis data.

Mesures	Results
P0	0.49
R20	0.29

Table 4. Results of KNN on Drug-related data.

Mesures	Results
P0	0.22
R20	0.07

Table 5. Results of the data-poor classifier on Diagnosis contents.

Mesures	Results
P0	**0.75**
R20	**0.63**

Table 6. Results of the combined system.

From a practical perspective, these quantitative results mean that: 1) the top ranked diagnosis is good three times out of four, and out of an average of five codes, the system can reliably predict three codes in the top twenty. This level of performance seems sufficient to improve encoding and billing with DRG (Diagnosis Related Group). From a comparative perspective and overall, in the current tuning of the system, results obtained by using the combined classifier are moderately superior to those obtained by using only the diagnosis fields, therefore we believe that using more powerful combination models will foster this trend: combining diagnosis fields together with the anamnesis and prescription should more radically improve the diagnosis prediction power of the application. From a medical point of view, it is interesting to observe that the anamnesis' content – when available – is sufficient to achieve a precision of more than 50% (P0 = 57) but it is less surprising if we consider that the anamnesis is available for patients, which are hospitalized more frequently than the average. These patients often suffer from chronic pathologies, which are repeated in the coding summary at each episode of care. The near 50% precision obtained by the prescription field is also expected considering that some medication can be very disease specific (e.g. *ventolin* for *asthma*).

Last but not least, it is also interesting to directly contrast these results with the task as performed by professional encoders. Thus, in our experiments, we assume that the coding provided by the institution's medical writers is a gold standard. However, the quality of the benchmark can be questioned. Indeed, it is well known that inter-coder agreement when measured on large controlled-vocabulary is generally significantly below 100%. Thus, Henderson *et* al. report on an agreement of 83% when only the main diagnosis is coded and 79% when all diagnosis codes are considered [31]. The number of expected categories is not reported in this study, but with 5.1 codes per case in our experiments, the task that we evaluate is likely to have a lower inter-coder agreement. This means that the theoretical upper bound effectiveness of our task is probably below 79%,l which suggests that a precision of 75% is probably very close to this theoretical upper limit.

Related studies and Discussion

As already mentioned in the introduction, medical document encoding is a fairly well studies problem in medical informatics. However direct comparison with existing experiments is difficult. Thus Pakho-

mov *et* al., have studied the automatic assignment of ICD (-9) codes using much larger datasets and report that more than 48% of all diagnosis fields can be classified with 98% precision and 98.3% recall using data-driven classification methods. However, the way they define the task is clearly simpler if we consider that the input to be categorize is much less ambiguous. They indeed classify short diagnosis contents rather than a merged set of paragraph collected across several documents. While our task is both a multi-document and a multi-class task, Pakhomov *et* al. report that 83% of the textual instances they attempt to classify have a unique ICD category, while 0.10% is to be assigned five ICD categories. In contrast, the proportion of episodes of care receiving a dozen or more than a dozen of codes is above 5% in our dataset.

In 2007, a Natural Language Processing challenge has been organized on the topic. Our group participated in the challenge (labelled "SIM" in the online result table). In a week of full-time equivalent work, we adapted both our data-poor and our data-intensive systems. Before submission, we compared the results obtained by each method but were unable to obtain a statistically significant difference between the two methods. We attempted to combine them but without much success. Finally, we decided to submit the data-poor run as official run, as it was performing 0.01% better than the other runs. As final result, our group's submission was ranked 17th out of more than 40 participating teams. Interestingly, there was no statistical difference between most of the 40 submitted runs. Basically, all runs submitted by all groups using various natural language processing and or learning methods, obtained near similar performances. Considering the task's theoretical upper bound (~80%) with the 90% precision achieved by top-performing participants, we consider that it is difficult to draw any conclusion out of this evaluation. We suggest that given the near equivalent performances obtained by the various methods proposed by the competitors, it is very likely that the marginal differences were due to over fitting phenomena. Finally, given the limited size of the provided data sets, as well as the limited space category set (25 ICD codes), used in these experiments, we believe the task should not be regarded as a realistic medical encoding task.

CONCLUSION

We have reported on the design and preliminary evaluation of an automatic text categorization engine to help medico-economic encoding in the University Hospitals of Geneva. The system obtains a precision at high ranks of 75% (P0) and a recall of 63% (R20). In addition to diagnosis contents, the direct use of diagnosis-related temporal events such as the anamnesis (i.e. the past) and the prescription (i.e. some future outcomes), results in a statistically significant improvement. This conclusion supports the importance of temporal text mining for medical decision support [16]. Current performances seem sufficient to improve encoding and case billing with DRG in the institution but further experiments will be needed to establish the impact of the system on the productivity and income of the hospital.

As future research direction, we would like to use the score returned by the system to assess the quality of the prediction. Indeed, we have shown elsewhere that the categorization status value (i.e. the score) returned by an ACT system can be used to suggest only high confidence categories [16]. Hypothetically, the system could be generic enough to be applied to several coding scenarii: from fully-automatic encoding tasks – by trading recall for precision – to interactive coding tasks as investigated in this paper. Finally, in the final application, we plan to provide for each predicted category a short passage to help the validation/rejection of the predicted codes by the professional encoder. As proposed in [16] to support functional annotation of proteins in the Swiss-Prot database, we believe such a reading assistant is necessary to speed up professional encoding, like it has been shown necessary to guide protein database curators.

dtu w_{ij} = $\frac{\left(\ln\left(\ln((tf_{ij}) * K_{Length(Feature)}) + 1\right) + 1\right) \cdot idf_j}{(1 - slope) \cdot pivot + slope \cdot nt_i}$
dtn w_{ij} = $[\ln(\ln(tf_{ij}) + 1) + 1] \cdot idf_j$

Table 7: Formula for dtu.dtn, modified to take into account the Length of the feature.

Acknowledgements

We would like to thank Claudine Bréant and Gilles Cohen for helping when accessing the warehouse.

References

[1] Lewis, D.D. Evaluating and Optimizing Autonomous Text Classification Systems. In Proceedings ACM-SIGIR'95, ACM Press, New York, 246-254

[2] A Aronson and O Bodenreider and H Chang and S Humphrey and J Mork and S Nelson and T Rindflesch and W Wilbur. The Indexing Initiative. A Report to the Board of Scientific Counselors of the Lister Hill National Center for Biomedical Communications, NLM, 1999.

[3] P Ruch, R Baud, and A Geissbühler. Learning-free Text Categorization, *AIME* 2003, LNCS/LNAI 2780, Dojat M; Keravnou E; Barahona P (Eds.).

[4] P Ruch. Query Translation by Text Categorization, *COLING 2004*, ACL Anthology, 2004.

[5] P Ruch. Automatic Assignment of Biomedical Categories: Toward a Generic Approach. *Bioinformatics*, 2006.

[6] LM de Bruijn and A Hasman and JW Arends, Automatic SNOMED classification - a corpus based method, Yearbook of Medical Informatics, J van Bemmel and AT MacCray, 1999.

[7] Pakhomov SV, Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. J Am Med Inform Assoc. 2006 Sep-Oct;13(5):516-25.

[8] Gay CW, Kayaalp M, Aronson AR. Semi-automatic indexing of full text biomedical articles. AMIA Annu Symp Proc. 2005;:271-5.

[9] Fox E.A. and Shaw J.A. (1994). Combination of multiple searches. In Proceedings TREC-2, (pp. 243-249). Gaithersburg: NIST Publication.

[10] Fujita S. (2004) Revisiting Again Document Length Hypotheses: TREC-2004 Genomics Track Experiments at Patolis. The Thirteenth Text Retrieval Conference, TREC-2004, Gaithersburg, MD.

[11] Alan R. Aronson, Dina Demner-Fushman, Susanne M. Humphrey, Jimmy Lin, Hongfang Liu, Patrick Ruch, Miguel E. Ruiz, Lawrence H. Smith, Lorraine K. Tanabe, W. John Wilbur (2005) Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. TREC Proceedings. TREC 2005, Gaithersburg, MD, USA.

[12] Christian Lovis, Robert H. Baud, Anne-Marie Rassinoux, P. A. Michel, Jean-Raoul Scherrer (2007) Building Medical Dictionaries for Patient Encoding Systems: A Methodology. AIME 1997: 373-380

[13]F Ehrler, A Geissbuhler, A Yepes, P Ruch (2005) Data-poor Categorization and Passage Retrieval for Gene Ontology Annotation in Swiss-Prot, *BMC Bioinformatics, vol 6*, Special Issue on BioCreative I.

[14] J Savoy, A Stemming Procedure and Stopword List For General French Corpora, Journal of the American Society for Information Science, p. 944-952, 50(10), 1999.

[15] Singhal A (2001) Modern Information Retrieval: A Brief Overview. IEEE Data Eng. Bull. 24. p 35-43

[16] Zhou L. and Hripcsak G. (2006) Temporal reasoning with medical data – A review with emphasis on medical language processing, Journal of Biomedical Informatics, 40. 183-202. 2007.

[17] P Zweigenbaum, R Baud, A Burgun, F Namer, E Jarrousse, N Grabar, P Ruch, F Le Duff, JF Forget, M Douyère and S Darmoni. UMLF: a unified medical lexicon for French. International Journal of Medical Informatics, 2004.

[18] Breant C, Thurler G, Borst F, Geissbuhler A. Design of a Multi Dimensional Database for the Archimed DataWarehouse, Stud Health Technol Inform. 2005;116:169-74

[19] C van Rijsbergen, Informational Retrieval, Butterworths, 1979.

[20] G Cohen, P Ruch, M Hilario: Model Selection for Support Vector Classifiers via Direct Simplex Search. FLAIRS Conference 2005: 431-435

[21] P Ruch and L Perret and J Savoy. Features Combination for Extracting Gene Functions from MEDLINE. 27th European Colloquium on Information Retrieval - *ECIR*, LNCS 3408, 2005.

[22] P Ruch, R. Baud, A.-M. Rassinoux, P. Bouillon and G. Robert 2000. Medical document anonymization with a semantic lexicon. In Proceedings of AMIA'2000, Los Angeles.

[23] Fabrizio Sebastiani: Machine learning in automated text categorization. ACM Comput. Surv. 34(1): 1-47 (2002)

[24] Singhal, A. (2001). Modern information retrieval: A brief overview. IEEE Data Eng. Bull., 24(4), 35–43.

[25] Manber, U. and Wu, S. (1994). GLIMPSE: A tool to search through entire file systems. In Proceedings of the USENIX Conference, pages 23–32, San Fransisco CA USA.

[26] Yang, Y. (1996b). Sampling strategies and learning efficiency in text categorization. In Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access.

[27] Wilbur, J. andYang, Y. (1996). Ananalysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. Comput. Biol. Med., 26(3), 209–222.

[28] Yang, Y. and Chute, C. (1992). A linear least squares fit mapping method for information retrieval from natural language texts. COLING, pages 447–453.

[29] Ruch P, Geissbühler A, Gobeill J, Lisacek F, Tbahriti I, Veuthey AL, Aronson AR. Using discourse analysis to improve text categorization in MEDLINE. Medinfo. 2007;12:710-5.

[30] Gobeill J, Tbahriti I, Ehrler F,Mottaz A, Veuthey AL and Ruch P. Gene Ontology density estimation and discourse analysis for automatic GeneRiF extraction. BMC Bioinformatics 2008, 9(Suppl 3):S9

[31] Henderson T, Shepheard J, Sundararajan V. Quality of diagnosis and procedure coding in ICD-10 administrative data. Med Care. 2006 Nov;44(11):1011-9.