Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230262

Automated ICF Coding of Rehabilitation Notes for Low-Resource Languages via Continual Training of Language Models

Kevin ROITERO[a], Andrea MARTINUZZI[b], Maria Teresa ARMELLIN[b], Gabriella PAPARELLA[b], Alberto MANIERO[b] and Vincenzo DELLA MEA[a,1]
[a] *Dept. of Mathematics, Computer Science and Physics, University of Udine, Italy*
[b] *Dept. of Conegliano, IRCCS "E. Medea" Scientific Institute, Conegliano, Italy*
ORCiD ID: Kevin Roitero https://orcid.org/0000-0002-9191-3280,
Andrea Martinuzzi https://orcid.org/0000-0002-0319-3579,
Vincenzo Della Mea https://orcid.org/0000-0002-0144-3802

Abstract. The coding of medical documents and in particular of rehabilitation notes using the International Classification of Functioning, Disability and Health (ICF) is a difficult task showing low agreement among experts. Such difficulty is mainly caused by the specific terminology that needs to be used for the task. In this paper, we address the task developing a model based on a large language model, BERT. By leveraging continual training of such a model using ICF textual descriptions, we are able to effectively encode rehabilitation notes expressed in Italian, an under-resourced language.

Keywords. Language Models, ICF, Rehabilitation, Continual Training

1. Introduction

The International Classification of Functioning, Disability and Health (ICF) is a classification of functioning conditions developed by the World Health Organization (WHO) [1]. The concept of functioning on which ICF is built upon is that of a "dynamic interaction between a person's health condition, environmental factors and personal factors". This overall model has been translated into a classification covering all the main components of functioning, namely Body Structures and Functions, Activities & Participation, and Environmental Factors.

Coding natural language (NL) notes with ICF is recognized as difficult task [2] and also with low inter-observer agreement [3]. Yet, the need for coding electronic health records also with ICF has been recognized as useful [4]. Thus, the use of tools to support this task is welcomed, although not yet researched as much as for other biomedical classifications like ICD, e.g., in [5,6,7,8].

Training models for ICF coding presents a main difficulty if compared with other biomedical classifications. In fact, while other terminologies and classifications include specialized terms and concepts (disease and procedure names, anatomy parts, etc.), ICF, in particular in its Activities & Participation and Environmental Factors components,

[1] Corresponding Author: Vincenzo Della Mea, E-mail: vincenzo.dellamea@uniud.it

give specific meaning within the ICF biopsychosocial framework to commonly used terms and concepts (e.g., Walking, Washing oneself, Food, Temperature, etc.). However, there have been seminal attempts to automated ICF coding ([9]), and recently attention resumed ([10,11]), also and notably in an under-resourced language like Dutch ([12]).

In the present paper, we propose a methodology for automated coding of NL texts using ICF, in an under-resourced language as Italian, thanks to the availability of a real-world dataset provided by a rehabilitation clinic in North-Eastern Italy. A preliminary pilot experiment probed the possibility of automatically code NL strings to the correct ICF code in 2 chapters of ICF: one from the Body Functions domain (functions related to movement) and one from the Activity & Participation domain (mobility).

2. Methods

2.1. Dataset

Recognizing the power of ICF in univocally describing the various aspects of human functioning, the use of codes from the whole ICF coupled with the appropriate qualifiers was systematically introduced from 2014 in a tertiary neurorehabilitation centre in Northern Italy. To ease initial use and to allow more accurate description of the content of the ICF entity used, each code is associated with a natural language free text where the professional describes in lay terms what he/she observes.

Since 2014, over 2000 complete functioning profiles of patients admitted to the centre have been stored and are now available for analysis. Health conditions span from cerebral palsy to Parkinson and patients age from 0 to > 80 years. These projects contain detailed descriptions attached to selected ICF codes of any level and the attached qualifiers. Retrievable data include: ICF codes and qualifiers used; Natural language description of the item; Linkage between ICF categories and evaluation tools.

Additional data for continual training was collected with the aim of focusing on ICF expressions. Since the kind of notes found in the rehabilitation dataset are not available from other sources, we identified some relatively suitable, yet sub-optimal, Italian texts in the ICF classification itself [1] and the Wikipedia article describing it.

The dataset was pre-processed as follows. We started with the full dataset and we removed the classes which appeared only once, which are 129; this left us with 473 classes that have associated at least two instances. Then, we subset the dataset in two different ways, as follows. To create the former dataset, we selected all the instances that belong either to the B7 or D4 chapters (associated to 107 classes). These classes are related to the important area of mobility from two different points of view: Body Functions (B7) and Activities & Participation (D4). Then, we split the dataset into the training and test sets performing stratified sampling; this left us with 6,650 instances in training and 1,751 in test. To create the latter dataset, we augmented the training set of the first one with instances not belonging to the B7 or D4 classes; the rationale behind this choice is that by leveraging more data the models we use can better tell apart the B7 and D4 classes, both each other and also from all the other classes. This second dataset share the test set with the former dataset and has 30, 438 instances in training.

2.2. Model training and inference

To develop and train our models we rely on both the PyTorch and HuggingFace frameworks. The models were trained and tested on a Linux server equipped with Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz, 64GB of RAM, and 2x Nvidia Geforce RTX 3090 GPUs. The trained models are available for research purposes.

BERT [13] is a transformer based model pre-trained on large corpora of multilingual text using a self-supervised training procedure. More in detail, the model has been trained on the textual corpora without the usage of any human annotation or feedback thus by relying on a sampling technique which is used to generate inputs and labels from the raw text. The model is trained with two objectives, Masked Language Modeling (MLM) and Next Sentence Prediction (NSP): the former receives in input a sentence and randomly masks a percentage (usually 15%) of the input tokens and asks the model to predict the masked tokens, while the latter takes two sentences in input which are either one after the other in the original text or not with a given probability (usually 50%) and asks the model to predict such relation between sentences (i.e., whether they were following each other in the input text or not). Note that such two training procedures are different from the classical ones used by recurrent networks which receive in input the words or tokens one after the other, and also from auto-regressive models like T5 [14], which receives mask only future tokens. The generated pre-trained model can be then fine-tuned on a variety of downstream tasks such as classification, regression, etc.

In this work, we rely on the *bert-base-multilingual-cased* model, a 110 million parameters model pre-trained on the largest Wikipedia dump in 104 different languages. Details on the training procedure are as follows: for the MLM training procedure 15% of the tokens have been masked; among those, in 80% of the cases the masked tokens are replaced with "[MASK]", in 10% of the cases thy are replaced with a random token, and in remaining 10% of the cases the input sentence is not altered. For NSP the probability for sampling subsequent sequences is set to 50%.

Starting from the base model, we developed two model variations; in the former we simply attach a classification head to the model to fine-tune it to our dataset, while for the latter we perform "continual training" before attaching the classification head: we continue the pre-training procedure of the model on ICF text for additional 100 epochs for the MLM objective keeping the sampling parameters as for the original model. We monitored the model losses to make sure the training procedure was stable and we did not encounter over-fitting or damage model weights. We did not use the NSP training procedure for the final model because we found it does not provide any increase in the effectiveness of the model; this is consistent with literature results, e.g., the training procedures implemented in the HuggingFace framework for the BERT based models.

3. Results

Table 1 shows the effectiveness of the proposed approach. The first three columns of the table detail the base model used (i.e., BERT), the training objective (i.e., either the usage of the plain pre-trained model or the model where we performed continual training), and the training dataset (i.e., the 6,650 instances with B7-D4 or the full dataset consisting of 30,438 instances), while the latter columns detail the values for the Accuracy and F1 scores, computed at different cutoffs; note that macro-average weights all classes equally independently of their frequency, as it is obtained by computing the metric independently

for each class and then taking the average, while micro average aggregated the contributions from the different classes to compute the metric, hence it is usually preferred in the multi-class scenario.

Table 1. Effectiveness of the proposed approach.

Base model	Training objective	Training dataset	Accuracy at 1	Accuracy at 3	Accuracy at 5	F1 micro	F1 macro	F1 weighted
BERT	pre-trained	B7-D4	.683	.864	.905	.683	.179	.607
BERT	continual	B7-D4	.709	.882	.925	.709	.226	.654
BERT	pre-trained	Full	.760	.910	.943	.760	.231	.718
BERT	continual	Full	.779	.917	.949	.779	.260	.746

As we can see from the table, both the training objective and the training dataset have an impact on the effectiveness scores. More in detail, as we can see by comparing respectively the first and second, and the third and fourth row of the table, the continual training objective increases the effectiveness scores; this is probably due because the further pre-training of the model allows to capture the context and semantics of ICF.

Furthermore, as we can see by comparing the first and last two pair or rows in the table, the training set employed has an effect of the effectiveness scores; in particular, the full dataset achieves higher effectiveness scores. This is probably due because the additional training instances allow the model to better tell apart the test classes, both one another as well as with all the other classes.

Overall, we can see that the model which achieves higher effectiveness scores is the one obtained by leveraging both continual training as well as the full training dataset.

4. Discussion

The results of this preliminary experiments seem to suggest the feasibility of automated ICF coding from text notes, at least for the specific subset of codes related to mobility. In the best case, Accuracy@1 is 0.779, which could be the performance of a totally automated system. A coding support system giving the coder three options among which to choose could attain a higher accuracy (0.917).

However, when looking at macro-averaged F1, the results suggest caution. The substantially lower values are given by the fact that ICF include a fair number of codes, many of which rarely used. So, performance is high on frequent codes only. Similar works in the past obtained higher F1 scores, but focusing on a smaller set of target codes. For example, in [11] two datasets with 13 and 16 ICF codes respectively were used; in [9] coding involved five codes; in [12] 4 main ICF categories were mentioned. In our dataset we removed codes used only once, but still the total number of used codes was 107, and this, with the selected models, decreased performance.

One limitation of the present work is the fact that the dataset, coming from a clinical setup, includes codes each provided by a single coder, thus with no measure of inter-observer agreement among human coders and, possibly, coding mistakes.

5. Conclusions

The preliminary experiment on ICF coding provides some feasibility evidence, with still some room for improvement for less represented codes.

Future works to overcome this limitation include:

- experimentation of zero- or few-shots learning methods to better cover the least used codes;
- identification of possible sources of text for continual training, which do not need to be coded in ICF, but still mentioning relevant concepts;
- if datasets in other languages become available, experimentation with large transformer-based multilingual models with multiple training languages and/or high generalization (i.e., zero- and few-shot learning) capabilities.

Acknowledgements

We thank for the support AM received from the Italian Ministry of Health (RC2022-23).

References

[1] World Health Organization. International Classification of Functioning, Disability and Health. WHO; 2001.
[2] Schuntermann MF. The implementation of the International Classification of Functioning, Disability and Health in Germany: experiences and problems. International Journal of Rehabilitation Research. 2005;28(2):93-102.
[3] Okochi J, Utsunomiya S, Takahashi T. Health measurement using the ICF: test-retest reliability study of ICF codes and qualifiers in geriatric care. Health and quality of life outcomes. 2005;3(1):1-13.
[4] Ustun TB, Chatterji S, Kostansjek N, Bickenbach J. WHO's ICF and functional status information in health records. Health Care Financ Rev. 2003;24(3):77-88.
[5] Gobeill J, Ruch P. Instance-based Learning for ICD10 Categorization. In: CEUR Workshop Proceedings 2125, CEUR-WS.org; 2018.
[6] Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H. Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)-Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study. JMIR Med Inform. 2019 Sep;7(3):e14830.
[7] Roitero K, Portelli B, Popescu MH, Della Mea V. DiLBERT: Cheap Embeddings for Disease Related Medical NLP. IEEE Access. 2021;9:159714-23.
[8] Falissard L, Morgand C, Ghosn W, Imbaud C, Bounebache K, Rey G. Neural Translation and Automated Recognition of ICD-10 Medical Entities From Natural Language: Model Development and Performance Assessment. JMIR Med Inform. 2022 Apr;10(4):e26353.
[9] Kukafka R, Bales ME, Burkhardt A, Friedman C. Human and automated coding of rehabilitation discharge summaries according to the International Classification of Functioning, Disability, and Health. J Am Med Inform Assoc. 2006;13(5):508-15.
[10] Newman-Griffis D, Fosler-Lussier E. Automated Coding of Under-Studied Medical Concept Domains: Linking Physical Activity Reports to the International Classification of Functioning, Disability, and Health. Front Digit Health. 2021 Mar;3.
[11] Newman-Griffis D, Maldonado JC, Ho PS, Sacco M, Silva RJ, Porcino J, et al. Linking Free Text Documentation of Functioning and Disability to the ICF With Natural Language Processing. Front Rehabil Sci. 2021 Nov;2.
[12] Meskers CGM, van der Veen S, Kim J, Meskers CJW, Smit QTS, Verkijk S, et al. Automated recognition of functioning, activity and participation in COVID-19 from electronic patient records by natural language processing: a proof- of- concept. Ann Med. 2022 Dec;54(1):235-43.
[13] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.
[14] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(140):1-67.