

A Whole-Person Function Dictionary for the Mobility, Self-Care and Domestic Life Domains: a Seedset Expansion Approach

Ayah Zirikly^{1,2} Bart Desmet¹ Julia Porcino¹
Jonathan Camacho Maldonado^{1*} Pei-Shu Ho^{1*} Rafael Jimenez Silva^{1*} Maryanne Sacco^{1*}

[1] Rehabilitation Medicine Dept., Clinical Center, National Institutes of Health
Bethesda, MD

[2] Department of Computer Science, Johns Hopkins University
Baltimore, MD

{ayah.zirikly, bart.desmet, julia.porcino,
jonathan.camacho, pei-shu.ho, rafael.jimenezsilva, maryanne.sacco}@nih.gov

Abstract

Whole-person functional limitations in the areas of mobility, self-care and domestic life affect a majority of individuals with disabilities. Detecting, recording and monitoring such limitations would benefit those individuals, as well as research on whole-person function and general public health. Dictionaries of linguistic expressions related to whole-person function would enable automated identification and extraction of relevant information. However, no such resources currently exist, due in part to a lack of standardized coding and their availability mainly in free text clinical notes. In this paper, we introduce dictionaries of whole-person function in the domains of mobility, self-care and domestic life, built and evaluated using a small set of manually annotated clinical notes, which provided a seedset that was expanded using a mix of lexical and deep learning approaches.

Keywords: disability, terminology, terminology expansion, whole-person function, mobility, self-care

1. Introduction

Functional limitations affect a large proportion of the world’s population, and it is estimated that over 1 billion people (WHO, 2011) live with some form of disability. In Europe, about 135 million people live with disability¹, and the community of people with disability in the United States is 61 million large (Centers for Disease Control and Prevention, 2020). Social insurance and security organizations, such as the Social Security Administration (SSA) in the United States, aim to support this population, for example by providing financial aid in the form of disability benefits. Given the scale of the problem, such organizations need to allocate considerable and rising resources, both financial and in terms of personnel, to review disability applications and provide an eligibility decision. These responsibilities are even more challenging when there are no diagnosis codes associated with an applicant’s medical evidence. As an example, a claim for a person who suffers from impairments due to cancer would usually be coded using the International Classification of Diseases (ICD), which eases the eligibility determination process. However, disability benefits can also be granted on the basis of whole-person function limitations which cannot be attributed to a single medical diagnosis. Whole-person functioning is assessed at the

activity and participation level and represents a transactive process of interactions between a person, their environment, and the nature and demands of the activities that the person wants to do, needs to do, or is expected to do to fulfill one’s roles. Deciding cases at the whole-person level is more challenging, due to the lack of use of standardized codes for whole-person function within the medical community, and the fact that their descriptions are heavily semantically dependent. For instance, a note by social worker may include the following: *the patient cannot bathe himself alone without the help of a family member, if available*. Such mentions imply that this person cannot perform the bathing task independently, and that help is not always available. It is important to note that the World Health Organization (WHO) developed the International Classification of Functioning, Disability, and Health (ICF) to address function-related information coding, but unfortunately its adoption in the clinical community is optional and not widespread, especially in the United States. On the other hand, some European countries, such as the United Kingdom, have seen more initiatives to adopt ICF coding (Gimeno and Lin, 2017).

Finding and coding whole-person function information in clinical text is a nascent field in the Natural Language Processing (NLP) community. Technological systems that could identify relevant information could improve capture by the clinician, and unlock it for research and disability determination purposes. This paper joins previous work such as (Desmet et al., 2020) that calls for more work in this area to build NLP so-

* Equal contribution

¹<https://www.euro.who.int/en/health-topics/Life-stages/disability-and-rehabilitation/data-and-statistics/facts-on-disability>

lutions that can perform information extraction using terminologies or machine learning models trained on annotated data. However, very few systems have been developed to address function and disability information extraction (Thieu et al., 2021; Newman-Griffis et al., 2019b; Newman-Griffis and Zirikly, 2018). In this work, we focus on building dictionaries of linguistic expressions such as terms, collocations and short phrases, which we will collectively refer to as terms, for the following whole-person function types: mobility, self-care, and domestic life, where these types account for 13.7% (the top disability limitation), 3.6%, and 6.3% of total disabilities². We start with a seedset extracted from datasets that were manually annotated by domain experts, consisting of clinical notes from two sources: the National Institutes of Health (NIH) and SSA. This seedset contains mentions for each of the domains of interest, in addition to their associated ICF codes that are also captured during annotation. We expand the seedset by clustering the expanded terms using a combination of lexical and deep learning techniques, where we provide both the mention and the suggested ICF code. The final set of terminologies is evaluated against manually annotated datasets for recall and is also filtered by our domain experts to contain only relevant terms with the correct ICF code. This paper also makes these terminologies available to the research community. We summarize our contributions as follows:

- To our knowledge, these are the first publicly available dictionaries of terms related to mobility, self-care, and domestic life, to help in extracting information relevant to disability.
- Our terminologies provide the associated ICF codes, where both term and ICF code have been verified by domain experts.
- We push for sharing and developing a common understanding and background about the studied domain in the medical-related communities.

2. Related Work

Whole-person function research in the NLP and bioinformatics communities is a recent field that has been attracting momentum in recent years. For instance, Newman-Griffis et al. (2019a) emphasize the importance and the significant impact of the problem on public health in their call for action for more research in the field. Many reasons exist for the slower advances in this field compared to the general clinical one, including the lack of standardized terms, data access, and the challenges associated with the language to describe such terms (Desmet et al., 2020).

There has been some work that introduced datasets annotated for whole-person function, along with

fine-grained descriptions of the annotation scheme, given the recency of the problem and the need for many domain experts to collaborate on establishing a scheme (Thieu et al., 2021). Other studies focused on the information extraction aspect of the problem and extracted whole-person function phrases and mentions in a similar way to other Named Entity Recognition (NER) tasks (Agaronnik et al., 2020; Newman-Griffis and Zirikly, 2018; Newman-Griffis et al., 2019b). Terminologies in the clinical NLP field have proven to be effective in many extraction tasks such as drug names (Lerner et al., 2020), disease (Chun et al., 2006), and symptoms (Silverman et al., 2021). Creating and expanding terminologies is an active area of NLP research as well. Fan et al. (2019), for example, discuss terminology augmentation of dietary supplements using word embeddings. In recent work by Newman-Griffis et al. (2021) in the area of ICF coding, the authors aim at linking mobility terms to their corresponding ICF codes using classification methods. However, they emphasize the importance of having domain-specific terminologies, since they could alleviate some of the challenges they faced and reduce the need for expensive manual annotation of documents by domain experts. Thus, building publicly available terminologies that can be used by the NLP community and other entities that handle data related to function, such as SSA’s disability determination program, is timely and necessary.

3. Approach

Our work to build terminologies for whole-person function-related domains aims to foster more interest and work in this field by the NLP community, and is also motivated by the lack of adoption of standardized ICF codes in clinical notes. We focus on two domains: i) mobility (MOB); and ii) self-care and domestic life (SCDL). We start with an example of how this information is reported in clinical notes.

Patient is **able to walk with the help of cane or another assistive device**. He reports that he **can take a bath and clean himself**.

The mention “able to walk with the help of cane or another assistive device” represents the patient’s mobility state, where *walk* is the action of the mobility mention. Similarly, the mention “can take a bath and clean himself” represents information related to SCDL, where *take a bath* and *clean himself* are the action terms for SCDL. In our work we focus on the action terms as opposed to the full mentions, since the former are the key candidates for building terminologies.

²<https://www.cdc.gov/ncbddd/disabilityandhealth/infographic-disability-impacts-all.html>

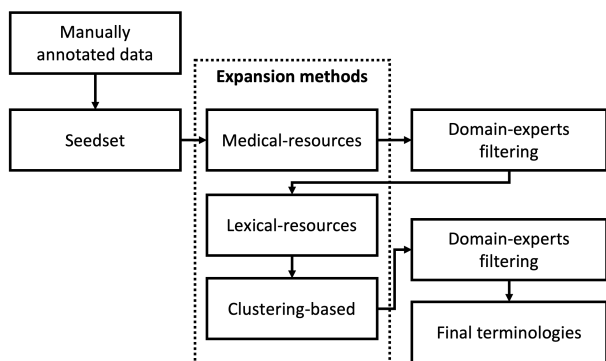


Figure 1: Approach summary

Figure 1 summarizes the steps we followed to generate expanded terminologies for both domains. In the following subsections we will cover each step of the process, which are identical for MOB and SCDL.

3.1. Seedset and dataset

The seedset for our terminologies is extracted from a dataset that was manually annotated by domain experts for both MOB and SCDL. The domain experts come from occupational therapy and public health backgrounds. Additionally, they were trained by physical therapists. The annotation is hierarchical: the full mention is annotated by its type (MOB or SCDL), and one or more actions and their corresponding ICF codes are annotated inside of those mentions. All the details related to the annotation process are described in Thieu et al. (2021).

The annotated dataset is a collection of notes from: (a) the NIH Clinical Center that was made available to our team through the Biomedical Translational Research Information System (BTRIS), providing 450 physical therapy notes for MOB and 30 for SCDL; and (b) the Social Security Administration (SSA), providing 250 Consultative Examination (CE) files annotated for both MOB and SCDL. CE files are generated by a medical source of an applicant that is requesting a disability eligibility determination. Each of the annotated datasets is split into a training and test set (80/20). The seedset is generated from the training set, and the test set is used for evaluation. We extract all the action items along with their ICF codes that are provided by the annotators (e.g. negotiate stairs [d400], swim [d455], self-bath [d510], change cloth [d540]). Then, we apply multiple text processing and normalization techniques:

- Manually correct Optical Character Recognition (OCR) errors. This step is mainly relevant to the SSA data, as CE files are mostly scans of medical documents that can be originally hand-written or printed.
- Tokenize the text using the spaCy tokenizer (Honnibal and Montani, 2017).
- Normalize text by removing numbers and some punctuation marks (e.g. periods or semicolons).

Our final seedset contained 426 MOB and 298 SCDL terms.

3.2. Expansion methods

In this section, we discuss the expansion methods we applied on our seedset.

3.2.1. Resource-based expansion

Medical resources Although whole-person function information is not well-coded and it is harder to find in medical resources, limited initiatives exist that created mappings between the ICF and the Systemized Nomenclature of Medicine (SNOMED) codes to improve functioning and disability coding (HHS, 2006). We use these mappings to extract SNOMED codes for the ICF codes that are in our seedset. Following that, we search the Unified Medical Language Resource UMLS (Bodenreider, 2004) for the mapped SNOMED codes and extract all the strings from the parent and the children nodes and add them to our dictionary. All the added terms inherit the same ICF code as the original term that it was expanded from.

Before moving to other expansion methods, we asked our domain experts to inspect the terminology list, filter out any irrelevant terms and correct the ICF codes if necessary. The reason for asking our experts at this stage, rather than later in the process, is the high confidence of the previous step and to make sure that the following expansion steps use only verified terms, in order to save our domain experts' time.

Lexical resources Given that action items typically contain verbs, VerbNet was an intuitive choice as our lexical resource (Schuler, 2005). We note that this choice means that for terms that are not verbs, VerbNet is not well-suited, and that it does not work well on multiword expressions such as “take bath”. In order to work around these obstacles, we generate expansions of the first word only and concatenate the rest of the term to the expansion candidate. Algorithm 1 describes the details of our expansion technique. Examples of

```

for  $term \in dictionary$  do
  clean and normalize  $term$ 
  if first word of  $term$  is verb then
    get all synonyms of  $term$ 
    add synonyms +  $term[1 : len(term)]$  to
    dictionary where ICF code ==  $term$ 's
    ICF code
  end
end
  
```

Algorithm 1: Expansion method using VerbNet

VerbNet expansions are listed in Table 1. For instance, for the term “ascend”, relevant candidate terms such as “climb” and “descend” are generated. However, terms such as “use” as an expansion for “run” is not relevant to the studied domain and will be eliminated during the expert’s filtering step.

Term	Verbnet expansion terms
run	fly, jog, jump, march, walk, use, play, ...
ascend	climb, descend, pass, bend, run, slant, ...
ride	oar, paddle, pedal, ride, row

Table 1: Examples from VerbNet expansion

We also explored the possibility of using BioVerbNet (Majewska et al., 2021), but after analyzing the results, we found that VerbNet was producing better results. We attribute this to the fact that most of the action terms are not medically oriented, which makes VerbNet a more adequate lexicon for this task.

3.2.2. Clustering-based expansion

For this technique we use the set of terms generated from 3.2.1 and generate word embeddings for each of the terms. We experimented with fastText (Bojanowski et al., 2016) and Word2Vec (Mikolov et al., 2013) on in-domain and out-of-domain data. Early experiments showed that fastText yields better performance. We assume this is because fastText’s character-level architecture is better able to handle idiosyncratic spellings and out-of-vocabulary terms. In the remainder, we use fastText embeddings only. For out-of-domain data we use Wiki word vectors trained on English Wikipedia (Bojanowski et al., 2017). For in-domain data we use BioWordVec (Zhang et al., 2019), which is a fastText model that was trained on PubMed data and on MIMIC-III - a collection of medical and health-related deidentified documents for over 40,000 patients (Johnson et al., 2016). Thirdly, we trained a fastText model on ten years of NIH BTRIS data.

For multi-word terms, we use the average vector of all the words of the term as its final vector. Following the embeddings generation step, we expand the seedset terms by clustering their embeddings and those from the fastText model. We use Nearest Neighbor (NN) clustering in high dimensional space with the open source package `annoy`³, an algorithm by Spotify that is optimized to build indexes once, which avoids rebuilding the tree for each term lookup. We use cosine similarity as a distance metric, set the number of trees to 150, and retain the 10 nearest neighbors.

Upon comparing a sample of the expanded terms generated with each of the models, we found that BioWordVec yields better results than fastText on BTRIS. We attribute this to the homogeneity of NIH data, where all medical experts follow a specific protocol. This makes the model less able to generalize to medical entities outside NIH, such as the data from SSA that includes documents and notes from across the United States. Compared to the out-of-domain Wiki model, both in-domain models yield more relevant terms.

Table 2 shows examples of expanded terms, where we see as with VerbNet there is a combination of relevant

and noisy terms. Unlike VerbNet, we see terms such as “ten-meter” as candidates for expansion. It is understandable why such terms are generated, given their co-occurrence with terms such as “walk”. Sample review helped us to perform an extra automatic cleaning step that excludes terms that include measurements before the domain experts performed their final review. We are aware that recent advances in contextual lan-

Term	Clustering expansion terms
ride	bike, horseback, ski, bicycle
pull	move, straighten, push, lift
walk	ascend/descend, stair climbing, ldcw, ten-meter

Table 2: Examples from clustering expansion

guage models such as BERT (Devlin et al., 2018) have shown superior performance in many NLP tasks such as Named Entity Recognition (Li et al., 2020). However, given the nature of our task we doubt that contextual and sentence embeddings would be advantageous, since we are embedding terms without any context.

3.3. Final set

The outputs from all the previous methods are combined for our domain experts’ final review of the terms and the suggested ICF codes. The set was split and distributed evenly among four domain experts. All of the annotators that were involved in annotating the original datasets from NIH and SSA participated in this task. There was no cross-checking given their extensive experience in the domain and since good inter-rater agreement was achieved for annotating the original seedset. During the review, domain experts were provided with the original term and ICF code alongside each of its expansions that inherited the same ICF code. The resulting dictionaries 1) contain terms and their ICF codes, 2) for the Mobility, Self-Care and Domestic Life domains, 3) were evaluated by domain experts and 4) are, to our knowledge, the first publicly available resources of this kind. The dictionaries are available online at <https://github.com/CC-RMD-EpiBio/terminologies>. There are no plans to release the clinical notes that served as source data for these dictionaries, given their protected nature.

4. Results and Evaluation

After expansion, cleaning and expert filtering, our final set contains 2413 terms for MOB and 1191 terms for SCDL. The experts’ evaluation process cut the number of terms that were generated by lexical resources by approximately 43%, whereas the expanded terms that were generated by clustering with the neural embeddings were reduced by approximately 20%. This suggests that the expansion approach based on embeddings provides less noisy and more relevant candidates.

We evaluated the final sets for both MOB and SCDL on the held-out test data described in Section 3.1. We

³<https://github.com/spotify/annoy>

use recall as an evaluation metric, in two forms: (i) exact string match between the annotated action terms for MOB or SCDL and one of our terminologies; and (ii) partial string match, where full credit is given if there is at least one word overlap between an action term in the annotated test data and the terminology. Since partial-match can be too permissive as an evaluation metric, we report both metrics for comparison.

In Figure 2, we display the results for MOB expansion terms using both metrics for the lexical approach, the clustering approach using embeddings, and the combined filtered list. The latter achieves an exact-match recall of 60%, and partial-match recall of 96%. For SCDL (not pictured), the combined list attains a recall of 40% on exact match and 88% on partial match.

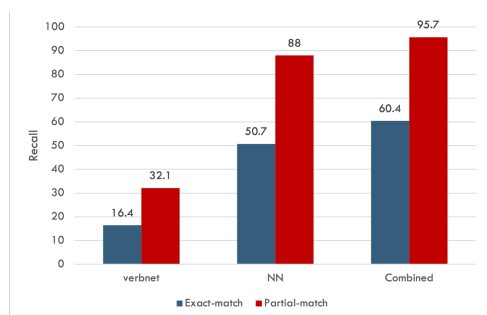


Figure 2: Evaluation results for MOB expansion term list: recall on held-out test data

Consistent with the filtering rate by the experts, the clustering approach on the embeddings yields superior performance for both partial and exact match. However, it is interesting to note that the combined version yields much better performance, which implies that the two approaches generate different types of candidates that are complementary. It is therefore suggested to pursue both approaches.

5. Conclusion

In this paper, we presented dictionaries of terms related to the whole-person function domains of Mobility (MOB), Self-Care and Domestic Life (SCDL), along with the appropriate codes from the WHO International Classification of Functioning, Disability and Health (ICF). We describe a terminology building approach based on manually annotated data, automatic expansion methods, and expert review, which resulted in curated lists of 2413 terms for MOB and 1191 for SCDL, that is available to the public. We find that a combination of approaches based on medical (SNOMED) and lexical (VerbNet) resources, and vector space nearest neighbor search on fastText embeddings, yields the best results.

Work on clinical coding, diagnoses and outcomes is abundant in the NLP community, and has achieved remarkable progress. Patients' functional status, however, is equally important to their wellbeing, independence and medical outcomes. Our hope is that this

work and resource facilitates more interest in, and research on, NLP approaches in the domain of whole-person function.

6. Bibliographical References

- Agarannik, N. D., Lindvall, C., El-Jawahri, A., He, W., and Iezzoni, L. I. (2020). Challenges of developing a natural language processing method with electronic health records to identify persons with chronic mobility disability. *Archives of Physical Medicine and Rehabilitation*, 101(10):1739–1746.
- Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.
- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.
- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Centers for Disease Control and Prevention. (2020). Disability Impacts All of Us. <https://www.cdc.gov/ncbddd/disabilityandhealth/infographic-disability-impacts-all.html>. [Online; accessed 15-Jan-2022].
- Chun, H.-W., Tsuruoka, Y., Kim, J.-D., Shiba, R., Nagata, N., Hishiki, T., and Tsujii, J. (2006). Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In *Bio-computing 2006*, pages 4–15. World Scientific.
- Desmet, B., Porcino, J., Zirikly, A., Newman-Griffis, D., Divita, G., and Rasch, E. (2020). Development of natural language processing tools to support determination of federal disability benefits in the us. In *Proceedings of the 1st Workshop on Language Technologies for Government and Public Administration (LT4Gov)*, pages 1–6.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arxiv. arXiv preprint arXiv:1810.04805*.
- Fan, Y., Pakhomov, S., McEwan, R., Zhao, W., Lindemann, E., and Zhang, R. (2019). Using word embeddings to expand terminology of dietary supplements on clinical notes. *JAMIA open*, 2(2):246–253.
- Gimeno, H. and Lin, J.-P. (2017). The international classification of functioning (icf) to evaluate deep brain stimulation neuromodulation in childhood dystonia-hyperkinesia informs future clinical & research priorities in a multidisciplinary model of care. *european journal of paediatric neurology*, 21(1):147–167.
- HHS. (2006). Disability public draft.
- Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings,

- convolutional neural networks and incremental parsing. To appear.
- Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9.
- Lerner, I., Paris, N., and Tannier, X. (2020). Terminologies augmented recurrent neural network model for clinical named entity recognition. *Journal of biomedical informatics*, 102:103356.
- Li, J., Sun, A., Han, J., and Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.
- Majewska, O., Collins, C., Baker, S., Björne, J., Brown, S. W., Korhonen, A., and Palmer, M. (2021). Bioverbnet: a large semantic-syntactic classification of verbs in biomedicine. *Journal of Biomedical Semantics*, 12(1):1–13.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Newman-Griffis, D. and Zirikly, A. (2018). Embedding transfer for low-resource medical named entity recognition: a case study on patient mobility. *arXiv preprint arXiv:1806.02814*.
- Newman-Griffis, D., Porcino, J., Zirikly, A., Thieu, T., Maldonado, J. C., Ho, P.-S., Ding, M., Chan, L., and Rasch, E. (2019a). Broadening horizons: the case for capturing function and the role of health informatics in its use. *BMC public health*, 19(1):1–13.
- Newman-Griffis, D., Zirikly, A., Divita, G., and Desmet, B. (2019b). Classifying the reported ability in clinical mobility descriptions. *arXiv preprint arXiv:1906.03348*.
- Newman-Griffis, D., Camacho Maldonado, J., Ho, P.-S., Sacco, M., Jimenez Silva, R., Porcino, J., and Chan, L. (2021). Linking free text documentation of functioning and disability to the ICF with natural language processing. *Frontiers in Rehabilitation Sciences*, page 67.
- Schuler, K. K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.
- Silverman, G. M., Sahoo, H. S., Ingraham, N. E., Lupei, M., Puskarich, M. A., Usher, M., Dries, J., Finzel, R. L., Murray, E., Sartori, J., et al. (2021). NLP methods for extraction of symptoms from unstructured data for use in prognostic COVID-19 analytic models. *Journal of Artificial Intelligence Research*, 72:429–474.
- Thieu, T., Maldonado, J. C., Ho, P.-S., Ding, M., Marr, A., Brandt, D., Newman-Griffis, D., Zirikly, A., Chan, L., and Rasch, E. (2021). A comprehensive study of mobility functioning information in clinical notes: entity hierarchy, corpus annotation, and sequence labeling. *International Journal of Medical Informatics*, 147:104351.
- WHO. (2011). World report on disability.
- Zhang, Y., Chen, Q., Yang, Z., Lin, H., and Lu, Z. (2019). Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):1–9.