Kickstarting NLP for whole-person activity with representation learning and data analysis

Denis Newman-Griffis[1,2] *Advisor:* Eric Fosler-Lussier[1] [1]The Ohio State University [2]National Institutes of Health, Clinical Center

[SRW - Thesis Proposal]

Background

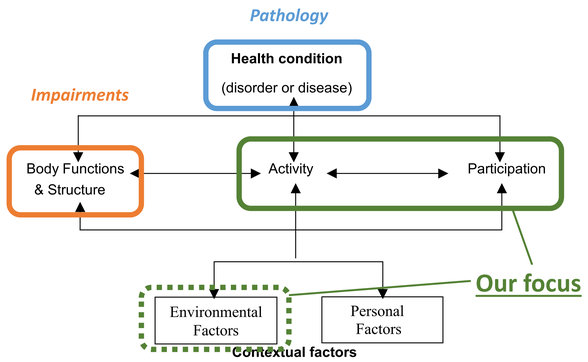

Figure 1. Diagram of human function, from International Classification of Functioning, Disability, and Health (ICF) (WHO 2001).

- Overall health and function can be conceptualized as an interaction between an individual, their environment, and actions they execute and roles they participate in.
- Clinical NLP works well for **pathology** (cell- and tissue-level) and **impairment** (organ- and system-level) information.
- Whole person information about **activities and participation** is a growing area of interest.
- Our focus: **activity reports** (information about activities and participation)
  - Highly relevant to disability determination
  - Complex descriptions with multiple components

> Pt slipped on icy **walk** and fell. Can now [**walk** by leaning on nearby objects, but only for ***short distances***].

Figure 2. Example of an **activity report**, illustrating (a) word ambiguity (**walk**), (b) non-standardized language (leaning on nearby objects), and (c) long-range dependencies (**short distances**).

Research Questions

RQ1 Activity reports exhibit linguistic characteristics distinct from other clinical language. **What kind of documents do activity reports appear in, and what structure do they have?**

RQ2 Many key terms and concepts for activity/participation and environment are not covered in existing resources. **How can representation learning address this coverage gap?**

RQ3 Activity reports often include common/ambiguous words. **How can representation learning support disambiguation in this domain?**

Characterizing rehabilitation language *RQ1*

- Rehabilitation medicine directly evaluates activity and participation
- Question: **How different are rehabilitation documents from other clinical records?**

Methods

- Analyze vocabulary frequencies in 3 corpora
  1. NIH Clinical Center (150k documents)
  2. OSU Wexner Medical Center (400k documents)
  3. MIMIC-III (2M documents)
- Use keyword frequencies to classify documents on three axes:
  - **Domain** – Functioning or Diagnostic information?
  - **Discipline** – Medical, Therapeutic, Psychosocial, or Administrative?
  - **Functional Area** – specialty within rehab (8-way)

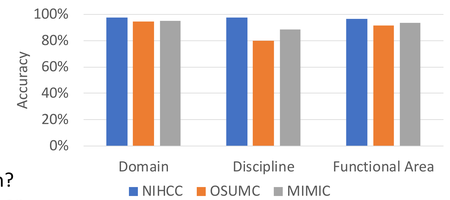

Figure 3. Classification accuracy using k-nearest neighbors with keyword frequencies, by axis and corpus.

Medical	Therapeutic	Psychosocial	Administrative
physical	supervision	past	take
surgical	independence	reports	tab
surgeon	admission	psychiatric	mouth
temperature	mobility	absolute	tabs
changes	therapy	care	date

Table 1. Top 5 keywords identified with labeled LDA for Discipline classes in OSUMC data.

Analyzing the structure of activity reports *(proposed work)* *RQ1*

Activity reports involve **interaction of multiple concepts**

- ✓ Individual
- ✓ Activity performed or role participated in
- ✓ Environmental factors (location, assistive devices, etc)

Action *Assistance*

The patient **ambulates** **with modified independence** for **300 ft**.

Quantification

Figure 4. A mobility activity report, broken down into constituent elements.

Question: **What elements define activity reports, and how are they connected?**

Methods

- Identify activities at different complexities
  - Walking, dressing, attending meals, going to work
- Curate list of key words/phrases for each (with domain experts)
- Find and filter hits for key terms in EHR data to find activity reports
- Identify atomic elements of each report
- Compare on three axes:
  1. Consistent dependency links between elements?
  2. Do frames align with existing frames in FrameNet?

No issue with 30-minute meetings with his manager during work

Talking action	
Independence	Independent
Location	Work
Duration	30-minute
Other party	Manager

The patient pushed exercise ball with her feet independently for 50 ft.

Move object	
Object	Exercise ball
Independence	Independent
Location	???
Distance	50 ft

Figure 5. Sample frames for two actions.

Representing concepts from unannotated data *RQ2*

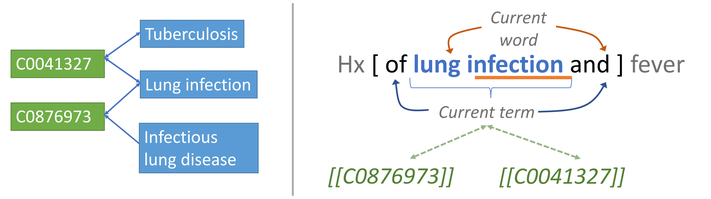

Figure 5. Example of a terminology mapping concepts to surface forms (left), and word, term, and distantly-supervised concept contexts (right).

- Corpora annotated for entity-level mentions are rare, esp. biomedical data
- We use terminologies for distant supervision
- We treat occurrences of a term as possible occurrences of each of its senses
- JET (Newman-Griffis et al, 2018): Jointly trains embedding models for words, terms, and entities

Pre-trained Wikipedia entity embeddings and UMLS embeddings from PubMed at https://slate.cse.ohio-state.edu/JET/

Normalizing Action types in activity reports *RQ3*

- We have developed dataset of 4000 activity reports (Thieu et al, 2017)
  - 3700 specific mobility-related actions, assigned one of 13 ICF codes; highly right-tailed distribution
- Normalizing these actions is challenging
  - Poor coverage in existing vocabularies
- We use JET and other methods to learn embeddings for ICF codes
- We train a DNN model to take embedding of Action context and candidate ICF codes and assign correct code

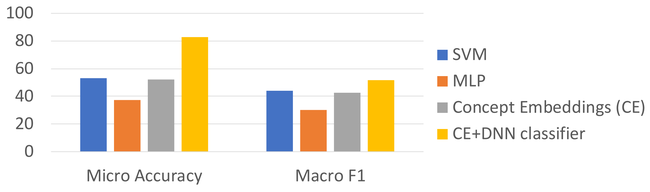

Figure 6. Accuracy and macro F1 for Action normalization, comparing projected concept embeddings to baseline methods.

Next steps

1. Analyze structure of activity reports in data from NIH and the US Social Security Administration
2. Investigate semi-supervised learning with Action normalization model for clinical concept normalization

Acknowledgments and References

1. International Classification of Functioning, Disability, and Health (ICF). World Health Organization: Geneva. 2001.
2. D Newman-Griffis, A M Lai, E Fosler-Lussier. "Jointly embedding entities and text with distant supervision." In *Repl4NLP*, 2018.
3. T Thieu, D Newman-Griffis, et al. "Inductive identification of functional status information and establishing a gold standard corpus: a case study on the mobility domain." In *BIBM*, 2017.

This research was supported in part by the Intramural Research Program of the National Institutes of Health, Clinical Research Center and through an Inter-Agency Agreement with the US Social Security Administration.