Diving for Pearls: Indexing Mobility Information in Social Security Administration Clinical Records with a Neural Relevance Tagger

Denis Newman-Griffis[1,2], Jonathan Camacho Maldonado[2], Pei-Shu Ho[2],
Albert M. Lai[3], and Eric Fosler-Lussier[1]

[1]Dept. of Computer Science and Engineering, The Ohio State University, Columbus, OH
[2]Rehabilitation Medicine Dept., National Institutes of Health Clinical Center, Bethesda, MD
[3]Institute for Informatics, Washington University in St. Louis, St. Louis, MO

Abstract

Locating sparse information in medical text that is relevant to self-reported functional limitations is a key challenge in the US Social Security Administration's process of determining disability. We investigate the effectiveness of a recent relevance scoring model for retrieving information related to mobility limitations, one of the most frequent allegation types in disability applications. Descriptions of mobility status are complex and difficult to extract with existing methods. We demonstrate that tagging for relevance at the token level achieves high recall on retrieving true mobility descriptions, and ranking documents by the amount of predicted mobility-relevant information achieves very strong correlation with ranking by the true number of mobility descriptions in each document. Additionally, experiments on a dataset of long, highly heterogenous documents show that our approach performs nearly perfectly at ranking documents with mobility-related information higher than those without, indicating that relevance estimation has high potential utility as a document triage tool for managing high-volume disability applications.

Introduction

The US Social Security Administration (SSA) is responsible for the management of the two largest federal disability programs in the United States, including the review and adjudication of new applications for disability benefits. The concept of "disability" is operationalized for SSA's purposes in terms of ability to meet the demands of gainful employment (US Social Security Administration, 2008). Some of the most frequent functional limitations leading to reduced ability to meet these demands relate to mobility activities, such as walking, transferring body positions, and using transportation (Courtney-Long et al., 2015). A key part of the disability adjudication process is the review of medical documentation to find evidence to support reported limitations. With current rates of disability applications placing high demand for adjudication (Autor, 2011), it is important to develop automated tools to assist with evidence finding in the adjudication process.

However, functional outcomes such as mobility are complex in nature, involving the interaction between an individual (including physiological aspects such as their health condition and their body structures as well as personal contexts), the action or role they wish to perform, and the environment they wish to perform it in (World Health Organization, 2001). As functional outcomes thus cannot be measured by pathology or physiological impairment alone (Altman, 2009), direct measurements are necessarily multi-faceted and, when recorded during medical encounters, typically as free text, lead to linguistically complex expressions. For example, the dataset presented by (Thieu et al., 2017) breaks descriptions of mobility status into multiple semantic components, including the action being performed, any source of assistance involved in

the performance, and one or more quantitative measurements of the event. Recent work on extracting these descriptions achieved reasonable token-level performance, but generalization of these findings is limited by the homogeneity of the corpus and the atypical nature of the patient population (clinical trial participants at the NIH Clinical Center) (Newman-Griffis & Zirikly, 2018). No comparable datasets of greater heterogeneity are yet available.

In this work, we investigate the use of a token-level neural relevance tagger to index mobility-related information in heterogeneous data associated with SSA disability applications. Mobility information is highly sparse in these documents, comprising on average less than 4% of document tokens (see Table 1). We evaluate the potential utility of our approach as an AI-assisted support tool for evidence review in disability adjudication, based on three specific use cases.

	CEs	1,200	
		CE	HIT
Num. documents	304	449	693
Annotation type	Spans	Document	Document
Avg. tokens/document	1,795.9	2,471.5	52,299.8
Avg. relevant segments/document	8.6	--	--
Avg. relevant tokens/document	70.7	--	--
Num. relevant documents	245	358	530
Num. irrelevant documents	59	91	163

Table 1. Details on two SSA datasets used for this study. Token count is given using SpaCy tokenization; for the span-level annotations, binary document relevance is defined as the presence of 1 or more relevant spans in the document. Span-level relevance statistics are not provided for documents in the 1,200-record corpus, as they are only annotated at the document level. 58 documents were removed from the 1,200-record dataset due to OCR noise.

Use case 1 is document review, evaluated in terms of token-level relevance tagging.

Use case 2 is fine-grained ranking of clinical documents by their expected amount of mobility information, evaluated in terms of ranking correlation.

Use case 3 is coarse-grained document triage to identify a high-impact set of documents for further analysis, evaluated on ranking relevant documents over irrelevant ones.

We demonstrate that a simple relevance tagging model yields strong performance on all three of these tasks, a first step in developing AI-based tools for reviewing functional status information. Qualitative review of system outputs shows complementary output patterns from static and contextualized embedding features, and identifies trends in output predictions and false negatives that suggest directions for further research.

The remainder of our paper is organized as follows. We first review related work on characterizing and identifying mobility descriptions in free text. We then describe the Social Security document sets we analyzed, followed by a description of our relevance tagging model and our experimental goals, with the relevant training and evaluation settings. We present our quantitative findings on token annotation and document ranking, and our qualitative results from analyzing tagger output from different experimental settings. Finally, we discuss the implications of our findings in terms of the potential utility of token-level relevance tagging for supporting evidence review during disability adjudication, and conclude with specific next steps for further research.

Related work

Functional status information is not well characterized in free text, but several prior studies have bearing on our work. Kuang et al. mine functional status-related terms from the web, and demonstrate that the majority of these terms are not covered in existing biomedical vocabularies

(Kuang et al., 2015). Skube et al. investigate functional status terms in post-surgical clinical notes, and identify several terms aligned with national quality indicators (Skube et al., 2018). Most recently, Newman-Griffis et al. investigate syntactic and semantic complexity of detecting the polarity of mobility descriptions (Newman-Griffis, Zirikly, Divita, & Desmet, 2019), and Doing-Harris et al. present a cardiac-centered vocabulary and ontology for frailty concepts, including several concepts related to functional status (Doing-Harris et al., 2019).

From the information retrieval (IR) perspective, relevance models have long been at the heart of IR research, and recent years have seen an uptick in neural network-based methods for IR; we refer the interested reader to (Onal et al., 2018) for an in-depth review. However, estimating relevance for functional status IR is relatively unexplored: Sundar et al. evaluate IR with structured functional activity codes (Sundar, Daumen, Conley, & Stone, 2008), and Kukafka et al. describe hand-crafted methods for assigning some such codes (Kukafka, Bales, Burkhardt, & Friedman, 2006), but studies from the text level directly are lacking. Our study thus brings neural IR techniques into the functional status domain, with concrete application to real-world SSA data.

Materials

We used two document collections for our study, both obtained from the US Social Security Administration through an Inter-Agency Agreement, and annotated by two domain experts; statistics of both document sets are provided in Table 1. The first consists of 304 consultative exams (CEs); these are special-purpose documents recording a detailed evaluation of the individual who filed the claim for disability benefits by an expert provider contracted by SSA for

the purpose (US Social Security Administration, 2014). Providers have typically not previously encountered the claimant, and these documents tend to be fairly long, but by and large consist of a set series of sections prescribed by SSA (US Social Security Administration, 2014). These documents were annotated for token-level span boundaries of mobility descriptions, following the protocol used by (Thieu et al., 2017).

The second document collection includes 1,200 documents drawn from two types of SSA records: additional CEs (disjoint from the first collection); and Health IT (HIT) documents, sets of records provided directly to SSA from provider EHR systems via regional Health Information Exchanges (HIEs) during the process of developing a disability case. Both of these document types were annotated with a binary label indicating the presence or absence of a substantive evaluation of mobility status anywhere in the document.

Two practical characteristics of the latter dataset impacted the annotation and analysis processes. Many documents were submitted to SSA via fax or scan, stored in image format, and converted into text documents using optical character recognition (OCR). As a result, the digital texts of many of these documents suffered from greater or lesser degrees of OCR noise. Our annotation protocol therefore included a provision that if a document was unreadable, or the status of mobility-related information in it could not be determined due to OCR or other noise, those documents would be removed from the dataset. After this filtering, our final dataset included 888 relevant documents and 254 irrelevant documents; further details are provided in Table 1.

Additionally, HIT documents in some cases consisted of a conglomeration of records from multiple encounters. Thus, each HIT document may include several individual records, sometimes spanning a significant time period. For the purposes of our annotations, we annotated the full document as relevant if any of the records it contained included mobility-relevant information. We did not include record segmentation or sectionization in our experiments, but highlight this as an important consideration for future work consuming HIE data where packaged records may not be automatically separated. This issue did noticeably increase both the size of our HIT documents (as shown in Table 1) and the sparsity of relevant information in them.

Methods

The linguistic characteristics of mobility descriptions are as yet poorly understood, and SSA data is unusually heterogeneous in both form and function, particularly compared to the homogeneous physical therapy notes used in the only prior study on mobility information processing (Newman-Griffis & Zirikly, 2018). Our focus in this study is on exploring the characteristics of mobility-related information in a setting where it is used in decision-making, and testing whether a simple approach to estimating relevance for information retrieval is is an effective support tool for triage of document sets. We experimented with HARE, a recent neural network-based method using word embedding features to estimate relevance at the token level (Newman-Griffis & Fosler-Lussier, 2019). Given both the novelty of mobility information and data privacy concerns pertinent to SSA data, we did not have access to well-developed baselines to compare against. However, our experimental goal is to evaluate AI-supported retrieval methods for what is currently a purely

manual review-based process, thus our hypotheses are evaluated purely in terms of recovering the target information at an acceptable level for decision support.

Relevance tagging model

The relevance tagger model we use, illustrated in Figure 1, processes a document token by token and assigns a relevance score between 0 and 1 with respect to the target information type (here, mobility). It takes as input word embedding features derived from the context around the current token, and passes these through a deep neural network with a binary softmax output layer.

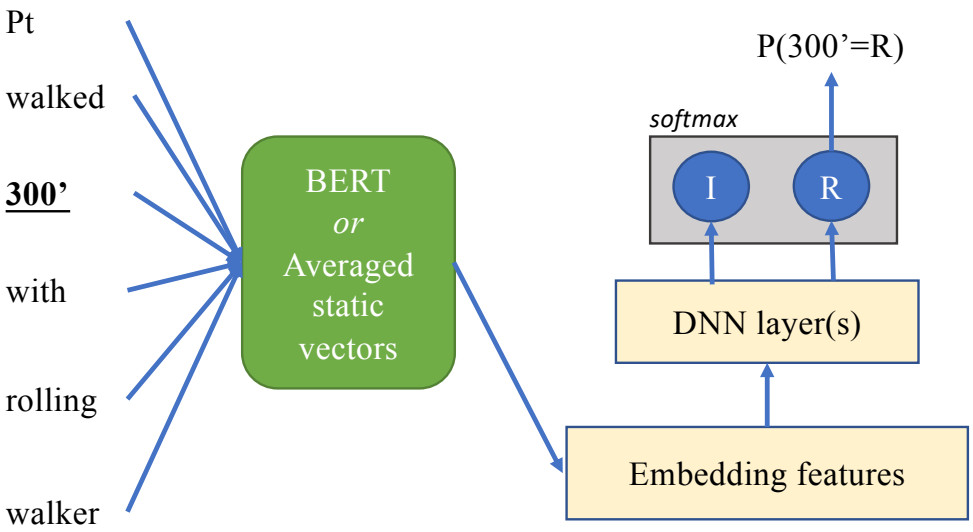

Figure 1. Relevance tagging model structure, tagging target token "300". For BERT features, the full input sequence is passed into BERT and the features at the target word index are passed into the DNN; for static features, embeddings are averaged across a fixed width window around the target word.

We experimented with two approaches to generate word embedding features: static and contextualized embeddings. For static embeddings, we utilized three 300-dimensional pretrained models: word2vec (Mikolov, Chen, Corrado, & Dean, 2013) trained on Google News,[1] GloVe (Pennington, Socher, & Manning, 2014) trained on 840 billion tokens of Common Crawl web

[1] Available from https://code.google.com/archive/p/word2vec/

text,[2] and FastText (Bojanowski, Grave, Joulin, & Mikolov, 2017) trained with subword information on combined Wikipedia and news data.[3] In addition, we trained our own FastText models on a separate corpus of approximately 70,000 medical evidence documents from SSA, using the skip-gram with negative sampling and CBOW training objectives; due to the much smaller corpus size, we trained 100-dimensional embeddings. Using static embeddings, we generated input features by averaging the embeddings for 10 tokens on either side of the target token (ending at linebreaks).

For contextualized embeddings, we used BERT (Devlin, Chang, Lee, & Toutanova, 2019), a language model-based model structure using a Transformer network to generate context-sensitive embedding vectors for each token in a sequence. We experimented with three pretrained BERT models: BERT-Base,[4] trained on Wikipedia and book data; BioBERT (Lee et al., 2019), trained on PubMed abstracts;[5] and clinicalBERT (Alsentzer et al., 2019), which is based on BioBERT but fine-tuned on clinical data.[6] All BERT models generate 768-dimensional vectors. We did not fine-tune the BERT models for our task, but rather generated embedding features using the fixed models and trained the HARE tagger on top of those features.

Training and hyperparameter settings

To train the HARE tagger, we used the token-level annotated data from the 304 CEs. These documents were tokenized by spaCy (Honnibal & Montani, 2017) (for static embeddings) or

[2] http://nlp.stanford.edu/data/glove.840B.300d.zip
[3] https://fasttext.cc/docs/en/english-vectors.html
[4] https://github.com/google-research/bert
[5] https://github.com/naver/biobert-pretrained
[6] https://github.com/EmilyAlsentzer/clinicalBERT

WordPiece (Wu et al., 2016) (required for BERT). We found that the document corpora we used did not lend themselves to clear definitions of sentence boundaries, and that the short segmentation often produced by clinical NLP toolkits (Griffis, Shivade, Fosler-Lussier, & Lai, 2016) frequently interrupted longer narratives; we therefore used linebreaks to separate text segments for embedding feature generation. We trained the model by subsampling a balanced set of relevant and irrelevant tokens from the full training set at each epoch, and training over this set of token samples using binary cross-entropy. After each epoch, we evaluated the model on a held-out 10% of the training data, and calculated F-2 score (which weights recall over precision), using a relevance score of 0.5 as the binarization threshold for discretizing model output. We used an early stopping threshold of 1e-05 on this development data, and trained with a patience of 5 epochs and a maximum training period of 50 epochs.

We experimented with the following hyperparameters of the relevance tagging model (results shown in Figure 2): number of hidden layers from 1 to 3, hidden layer dimensionality in {10, 100, 300, 768}, and per-layer dropout rate for noise-robustness from 0% to 90% (at 10% intervals). To evaluate different methods for controlling the imbalance between relevant and irrelevant tokens, we experimented with a positive fraction subsampling ratio per epoch from 10% to 100% (at 10% intervals), positive to negative sampling ratio in {0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, 3}, and positive to negative class weighting ratio in {1:1, 2:1, 3:1, 4:1, 5:1}.

To identify the best embedding models and hyperparameters for our experiments, we first compared development F-2 for each candidate model for static and BERT embeddings. BioBERT yielded the best contextualized performance; for static embeddings, word2vec GoogleNews

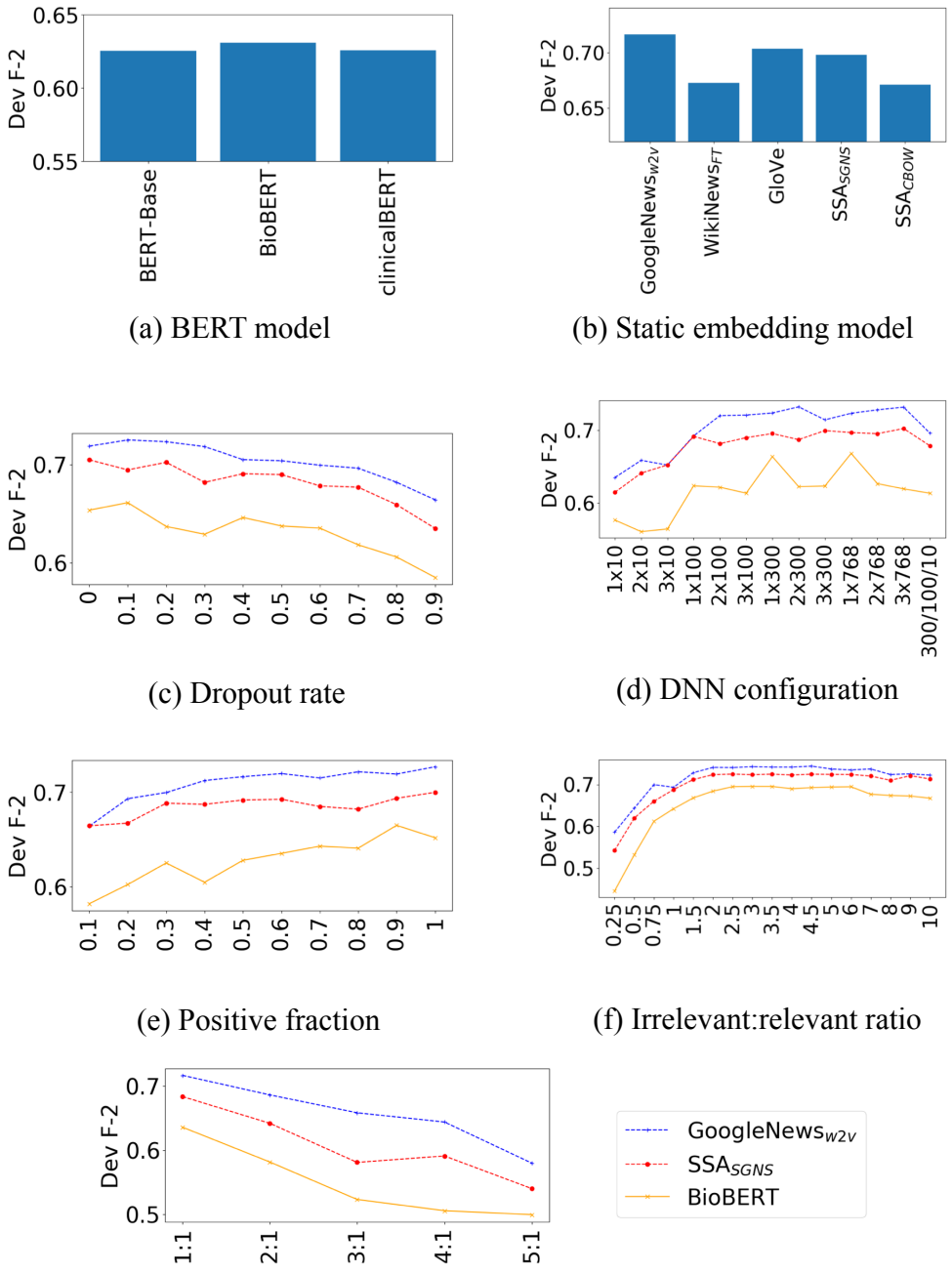

Figure 2. Cross validation development F-2 on 304 CEs for hyperparameter tuning experiments.

embeddings (GoogleNews$_{w2v}$) performed best of the out-of-domain embeddings, while skip-gram FastText embeddings from SSA data (SSA$_{SGNS}$) nearly matched its performance. In order to compare both in-domain and out-of-domain embeddings, we compare both SSA$_{SGNS}$ and

GoogleNews$_{w2v}$ embedding features, in addition to BioBERT. Averaged across these three sets of embedding features, our best performance on development data in cross-validation came from using a 1-layer 768-unit DNN with 10% dropout, and training at each epoch with all positive samples, negative samples at a 3:1 ratio, and equal class weights.

Experiments

We evaluated the utility of neural relevance tagging for three applications, representing different components of a clinical records review workflow. The most direct application (Experiment 1) is document review in order to locate evidence. Experiments 2 and 3 investigate information retrieval applications: Experiment 2 evaluates detailed ranking of different levels of mobility information, and Experiment 3 evaluates a purely triage application of ranking documents with any mobility information over those without any. These sets of experiments are discussed in detail in the following sections.

Experiment 1: token-level relevance tagging

Our first set of experiments was designed to evaluate the accuracy of our relevance tagger at the token level, as a strict measure of our ability to exactly recover the location of mobility-relevant information in SSA documents. Five-fold cross validation was used on the 304 CE corpus for these experiments; held-out development data for halting model training was randomly subsampled from the training set of each fold. At test time, all tokens of each test document were passed as input to the model, and the output relevance probability recorded for each. Evaluation

was conducted by binarizing the relevance probabilities at 0.5, and calculating precision, recall, F-1, and F-2 over the full set of test tokens.

Features	P	R	F1	F2
$GoogleNews_{w2v}$	48.9	82.2	61.2	72.3
SSA_{SGNS}	47.9	82.2	60.5	71.8
BioBERT	46.9	76.9	58.2	68.0

Table 2. Token-level relevance tagging results on 304 CEs from 5-fold cross validation, using each embedding model. Statistics are averaged across folds, using a relevance score binarization threshold of 0.5. P=Precision, R=Recall.

Table 2 shows the results from our three embedding methods. Relevance tagging at the token level achieves high recall in all three cases, although only roughly one in two tokens tagged by the model are "true" relevant tokens. Interestingly, static embeddings outperform contextualized BERT embeddings on both precision (1-2%) and recall (5.3%), suggesting that either the contextualized features are overparameterized for this size of dataset, or that using static embeddings enables leveraging lexicalized triggers in a way that the BERT model has not been tuned to do. For static embeddings, the higher-data $GoogleNews_{w2v}$ embeddings slightly outperform in-domain SSA_{SGNS} features (0.5% F-2). From an application perspective, however, all three embedding methods are effectively equivalent, providing high-recall indexing with a signal to noise ratio of about 1:1.

Experiment 2: document ranking

Our token-level tagging experiments measured our system's ability to strictly recover the information of interest. As highlighted in the Introduction, another application of use to SSA in processing large collections of medical evidence is priority ranking documents by the amount of mobility-related information they are likely to contain. We therefore conducted document ranking experiments, again using the 304 token-level annotated CEs. A gold standard ranking was

calculated by counting the number of relevant segments (i.e., contiguous sequences of relevant tokens) in the gold annotations for each document; in the case of a tie in number of segments, the document with the greater overall number of relevant tokens was assigned the higher ranking. The same ranking procedure was applied to binarized token-level relevance predictions to produce a model ranking.

Evaluation of our relevance tagging system was conducted using five-fold cross validation to obtain relevance scores for every token in the 304 CEs dataset in a test scenario, as in our first set of experiments. However, measuring rankings of five different 60-document sets is less informative for a high-volume scenario than ranking all 304 documents; we therefore combined the test set predictions from all five folds and ranked the full document set based on these. In our view, the practical evaluation at a larger scale outweighed potential cross-contamination effects of using test set outputs trained on overlapping training sets; nonetheless, this evaluation is necessarily somewhat optimistic. Ranking performance was measured using Spearman's rank correlation coefficient ρ, which ranges from -1 (indicating perfect anti-correlation) to +1 (indicating perfect correlation), where 0 indicates no correlation (i.e., random re-ranking). The relevance scoring toolkit of (Newman-Griffis & Fosler-Lussier, 2019) includes a Viterbi decoding-based smoothing technique for reducing noise in output relevance scores, which we include in our experiments.

Results As shown in Table 3, raw token-level relevance scores rank the 304 documents with very strong correlation to the gold ranking (ρ=.819 in the worst case). Viterbi smoothing increases ranking quality considerably, to ρ =.892 in the best case, without noticeably degrading the token-

	Raw		Smoothed	
Features	ρ	F-2	ρ	F-2
$GoogleNews_{w2v}$	.832	72.3	.887	72.1
SSA_{SGNS}	.826	71.9	.892	71.4
BioBERT	.819	68.1	.873	69.3

Table 3. Spearman's ρ and token-level F-2 for document ranking experiments on 304 CEs, evaluating on combined test set predictions from all folds of cross validation. F-2 is micro-averaged, slightly increasing over the macro-averaged F-2 in Table 2. Raw uses token-level relevance scores without post-processing; Smoothed includes Viterbi smoothing.

level annotation quality. The effect of smoothing on ranking correlation is about the same for all embedding models; however, its effect on token-level annotations is noticeably stronger when using BERT embeddings, where precision is increased by nearly 10% (to 56.2%; compared to a 3% gain for each of the static models), with a 4% drop in recall (3% for static models). Overall, all three models yield extremely strong correlation between model ranking and gold ranking when smoothing is applied, indicating that while token-level annotation may be noisy, it nonetheless captures the relevant information for successful retrieval.

Experiment 3: binary document ranking

At sufficient scale, determining whether a document merits further detailed analysis is an important first step in document triage and prioritization. Additionally, minor re-rankings of documents with similar amounts of mobility information may affect Spearman's ρ while having minor practical impact on system utility. We therefore conducted a third set of experiments evaluating document ranking based on a binary assessment of whether they were likely to have any mobility information in them or not. In this scenario, documents were ranked using the same procedure as in Experiment 2, and this ranking was compared to the gold document-level binary labels to report average precision (AP). AP measures, for each relevant document, the proportion

of the documents ranked higher which are truly relevant, and averages these ratios to report overall ranking quality.

For these experiments, we trained our relevance tagger using the full set of all 304 token-annotated CEs, and generated relevance scores for all tokens in each of the 1,200 binary-annotated documents. Due to the length of the HIT documents in this collection (an average of 52,000 tokens in each document), feature generation using BERT proved logistically infeasible: feature extraction on a subset of 150 documents took several days and produced hundreds of gigabytes of output. We therefore constrained our experiments to static embedding features only; results from our first two sets of experiments suggest that BERT would achieve comparable performance absent logistical difficulties.

Results Table 4 shows the average precision achieved for the 1,200 document dataset, overall and by document type. Both sets of static embeddings overwhelmingly rank relevant documents higher than irrelevant documents, achieving 97.1% overall AP in the worst case. As can be expected from the considerably larger size of HIT documents, they are slightly more difficult to rank correctly than CE documents are, though both feature sets yield above 97.5% AP. Thus, token-level relevance tagging is clearly effective for prioritizing relevant documents in a triage setting.

Features	CE	HIT	All
$\text{GoogleNews}_{\text{w2v}}$	48.9	82.2	61.2
SSA_{SGNS}	47.9	82.2	60.5
BioBERT	46.9	76.9	58.2

Table 4. Average precision for rankings of documents in our 1,200-document corpus, evaluated on binary document-level relevance annotations. Results are also broken out for CEs (449 documents) and HIT (693). BERT features were not used due to document length. Note that as CE and HIT documents are interleaved in the All setting, overall results can be lower than on individual subsets.

Qualitative analyses

Our quantitative system evaluations measured the utility of neural relevance tagging for different application scenarios. We also conducted qualitative analysis of system outputs to gain an understanding of what kinds of data are being tagged as relevant (correctly or erroneously) by different systems, and what implications these trends have for practical evidence retrieval of mobility-related information. We investigated three primary questions:

1. What differences do we observe in system outputs when using static vs contextualized word embedding features?
2. What patterns of error do we observe for false negatives, i.e. true mobility descriptions missed by our relevance tagger? (This analysis is constrained to the 304 CEs, as it requires token-level gold relevance annotations).
3. What patterns do we observe in text segments tagged as relevant? This includes both true and false positives in the token-level 304 CEs dataset, but also review of relevance annotations produced for the 1,200 document dataset.

Static vs contextualized features

While BERT and static embeddings yield comparable results in our experimental evaluations, they exhibit distinct patterns in the relevance annotations they produce. As shown in Table 5, BERT features lead to a striking increase in number of relevant segments tagged compared to static embedding features, and a concomitant reduction in the length of each segment. Many of these short segments are in fact close to one another, and are often parts of a longer segment tagged by static features; for example, BERT highlights the underlined phrases in the true segment

Features	# Segments		Tokens/Segment	
	Mean	Max	Mean	Max
$GoogleNews_{w2v}$	10.7	57	11.2	114
+Smoothing	6.7	42	16.3	130
SSA_{SGNS}	10.9	64	11.2	99
+Smoothing	6.4	39	17.3	103
BioBERT	46.5	319	2.8	71
+Smoothing	15.8	87	6.6	93

Table 5. Number of segments per document and number of tokens per segment, using relevance scores on the 304 CEs. Results are given using raw scores, binarized at 0.5, and with Viterbi smoothing. $GoogleNews_{w2v}$ and SSA_{SGNS} use SpaCy tokenization, while BioBERT uses WordPiece.

"her husband estimated that the maximum weight she could lift would be equivalent} to a gal ##lon of milk"; static features tag the entire segment contiguously. Many BERT segments are one or two-word phrases that appear somewhat random: for example, the underlined phrases in "her hair was brown and neck length". Interestingly, changing the binarization threshold does not noticeably decrease this noise without removing a considerable degree of useful signal as well. However, as illustrated in Table 5, Viterbi smoothing does close some of the gaps between segments and remove noisy segments, considerably decreasing the number of segments and increasing mean segment length. False positives remaining after smoothing are typically reasonable, if not necessarily directly relevant to mobility: for example, "her posture was within normal limits", and "she did not use correct ##ive lenses".

Static embeddings produce many fewer short segments, though some individual words and phrases are still tagged: e.g., "The claimant reported he has a problem with agitation". Static output segments, by contrast, often start before a true relevant segment and extend after it, suggesting lexicalization effects within the 10-token context window; see "enabled him to *take a*

job as a school bus driver" (italics indicate the true relevant segment). Some static segments are also offset from true segments, e.g. "*will get up once during the night* to use the bathroom". This produces an error in token-level evaluation, but is still helpful from a retrieval standpoint.

False negatives

In terms of false negative segments (i.e., true relevant segments in which none of the tokens were tagged as relevant), the noisiness of BERT output proves useful: virtually no relevant segments in the 304 CEs were entirely missed when using BERT features. Static embeddings, while qualitatively appealing in producing long, contiguous relevant segments, are also more susceptible to false negatives. The main trend we observed in these cases was syntactic: examples with mobility-relevant action verbs, such as "walking" or "transfer" are retrieved reliably, but many segments with mobility-relevant nouns were missed when using static features. For example, "right and left lateral bending approximately 10 degrees" was missed by static features, while BERT tagged "and" and "bending approximately." Some examples combining long-distance dependencies with less direct assertions, such as "claimant has symptomatic limitations in his ability to squat", were also missed by both sets of static features.

Relevance prediction patterns

Apart from the distinctions between outputs from static and BERT features, we observed several general patterns in relevance tagging outputs. The first is lexicalization: action verbs such as "stoop," "crouch," "climb," and "balance" (along with morphological variants) were tagged as relevant more than 90% of the time by all three embedding models, as were mobility-relevant

objects such as “ladders,” “ramps,” and “stairs.” These lexicalizations mostly reflected the true mobility annotations, though some false positives leaked through: for example, lexicalizing “table” from uses as a physical location led to 64 HIT documents being tagged with a single relevant segment, “Table of Contents.” More practically, lexicalization effects yielded some acceptable false positives, such as “`she left because she had surgery and had to stand`”, as well as more neutral statements such as “`call bell and possessions in reach`” (where “reach” is a common verb in mobility annotations).

The more challenging problem to handle in terms of false positives is a pragmatic one. Some phrases tagged as relevant describe neutral positional information: e.g., “`Patient in chair`”, with no information on how the patient reached that position. More significantly, many tagged references to mobility status were hypothetical, describing goals for the patient's therapy or conjectures; several others were field headers in clinical templates, referring to limitations or actions that may or may not have been observed. Local context is insufficient to capture these pragmatic implications in the absence of prior knowledge about template format or current section; finding systematic ways of incorporating this knowledge into relevance estimation systems represents a significant area for further work on improving AI-assisted tools for evidence review.

Features	# Segments	Tokens/Segment	% Tokens
CE (2,471.5 average tokens)			
$\text{GoogleNews}_{\text{w2v}}$	14.7	8.9	5.3
SSA_{SGNS}	14.9	9.0	5.4
HIT (52,299.8 average tokens)			
$\text{GoogleNews}_{\text{w2v}}$	140.3	6.5	1.7
SSA_{SGNS}	118.6	6.9	1.6

Table 6. Mean number of segments per document and number of tokens per segment, with the mean proportion of the tokens in each document tagged as relevant, for the CE and HIT subsets of the 1,200 documents. Results are given using raw scores, binarized at 0.5, without Viterbi smoothing.

Finally, as our document sets varied considerably in length, we investigated how relevance annotation scales to larger documents. Table 6 shows statistics for relevance annotations of the 1,200 document dataset. While the number of segments annotated as relevant increases on average for the longer HIT documents, the fraction of document text annotated as relevant decreases considerably, demonstrating that the relevance tagging model successfully ignores much of the irrelevant data introduced in the longer documents.

Discussion and limitations

Our experiments, while yielding compelling results, are simulations of real document review workflows. In order to evaluate potential utility for operationalizing AI-assisted tools such as relevance tagging within real-world disability adjudication at SSA, two clear next steps are needed. First, this work can be extended to other types of functional status information, either by developing multiple expert relevance taggers (e.g., one for mobility, another for domestic life activities, etc) and combining their output, or developing multi-stage models to gradually zero in on specific information of interest for an individual's claim. Second, a usability study and/or randomized controlled study should be conducted to evaluate whether disability adjudicators find integrating AI-assisted tools into the adjudication process helpful, and whether this integration results in meaningful process improvements.

Outside of the SSA setting, the potential utility of research on retrieving functional status information like mobility is limited by a lack of appropriate data. The SSA records used in this

study were heterogeneous in length, contents, source provider, and document type, but are subject to stringent data privacy protections. Similar protections apply to US Department of Veteran's Affairs data, another data source supporting valuable research in clinical outcomes (Shao et al., 2016). More accessible data sources, such as MIMIC (Johnson et al., 2016), are either from a single institution (and often a single specialty) or lack data relevant to functional status. Efforts to develop more diverse and accessible data sets about functional status will significantly contribute to research such as ours by facilitating easier comparison of systems and enabling a broader body of researchers to be involved.

A few limitations in our experimental evaluations should also be discussed. While our study was intended as a proof of concept for supporting evidence review with NLP, and data privacy and a lack of appropriate models made identification and use of relevant baseline methods difficult, it is quite possible that other methods would yield superior results for any of our three experiments. We therefore cannot make claims beyond *potential* utility, but by using a published method with publicly-available embedding features, our results can at least serve as a baseline for further research on evaluation and system comparison in similar settings.

Finally, two specific characteristics of the SSA documents we used affected our experimental results. Our filtering process of the OCRed portions of our documents only removed documents that were unreadable, leaving many documents with small amounts of remaining OCR noise; in all cases, line breaks were also introduced by the OCR process. OCR errors impacted model performance slightly, leading to "relevant" tags such as "`1-6-4;e2aw`" that were included in relevance evaluation. Additionally, we did not distinguish in our document set between documents

associated with different disability claimants; in a practical setting, document ranking would be applied to triage the records for a specific individual, which might affect the ranking quality.

Conclusion

We have demonstrated successful application of a neural relevance tagging model to support both document triage and evidence review related to mobility status in clinical records. Our results indicate the potential utility of AI-assisted technologies to support tasks involving a high volume of clinical data, such as disability adjudication. The effectiveness of a simple neural model using embedding features for our three sets of experiments establishes a strong baseline for new methodological research in retrieving functional status information, and highlights the value of an in-depth usability study with process experts to evaluate the impact of new technologies on the adjudication process.

This research was supported by the Intramural Research Program of the National Institutes of Health and the US Social Security Administration.

References

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., & McDermott, M. (2019). Publicly Available Clinical BERT Embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78. Retrieved from https://www.aclweb.org/anthology/W19-1909

Altman, B. M. (2009). Population survey measures of functioning: strengths and weaknesses. In *Improving the Measurement of Late-Life Disability in Population Surveys: Beyond ADLs and IADLs: Summary of a Workshop* (pp. 99–156). Washington, DC: The National Academies Press.

Autor, D. H. (2011). *The Unsustainable Rise of the Disability Rolls in the United States: Causes, Consequences, and Policy Options*. https://doi.org/10.3386/w17697

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the ACL*, *5*, 135–146. https://doi.org/1511.09249v1

Courtney-Long, E. A., Carroll, D. D., Zhang, Q. C., Stevens, A. C., Griffin-Blake, S., Armour, B. S., & Campbell, V. A. (2015). Prevalence of Disability and Disability Type Among Adults--United States, 2013. *MMWR. Morbidity and Mortality Weekly Report*, *64*(29), 777–783. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26225475

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Retrieved from https://www.aclweb.org/anthology/N19-1423

Doing-Harris, K., Bray, B. E., Thackeray, A., Shah, R. U., Shao, Y., Cheng, Y., … Weir, C. (2019). Development of a cardiac-centered frailty ontology. *Journal of Biomedical Semantics*, *10*(1), 3. https://doi.org/10.1186/s13326-019-0195-3

Griffis, D., Shivade, C., Fosler-Lussier, E., & Lai, A. M. (2016). A Quantitative and Qualitative Evaluation of Sentence Boundary Detection for the Clinical Domain. *AMIA Summits on Translational Science Proceedings 2016*, 88–97. American Medical Informatics Association.

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To Appear*.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., … Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*, 160035. https://doi.org/10.1038/sdata.2016.35

Kuang, J., Mohanty, A. F., Rashmi, V. H., Weir, C. R., Bray, B. E., & Zeng-Treitler, Q. (2015). Representation of Functional Status Concepts from Clinical Documents and Social Media Sources by Standard Terminologies. *AMIA Annual Symposium Proceedings 2015*, 795–803.

Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765559/

Kukafka, R., Bales, M. E., Burkhardt, A., & Friedman, C. (2006). Human and Automated Coding of Rehabilitation Discharge Summaries According to the International Classification of Functioning, Disability, and Health. *Journal of the American Medical Informatics Association*, *13*(5), 508–515. https://doi.org/10.1197/jamia.M2107

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *ArXiv Preprint ArXiv:1901.08746*, 1–8. Retrieved from http://arxiv.org/abs/1901.08746

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space [Computation and Language]. *ArXiv Preprint ArXiv:1301.3781*, 1–12. Retrieved from http://arxiv.org/abs/1301.3781

Newman-Griffis, D., & Fosler-Lussier, E. (2019). HARE: a Flexible Highlighting Annotator for Ranking and Exploration. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, 85–90. Retrieved from https://www.aclweb.org/anthology/D19-3015

Newman-Griffis, D., & Zirikly, A. (2018). Embedding Transfer for Low-Resource Medical Named Entity Recognition: A Case Study on Patient Mobility. *Proceedings of the BioNLP 2018 Workshop*, 1–11. Retrieved from http://aclweb.org/anthology/W18-2301

Newman-Griffis, D., Zirikly, A., Divita, G., & Desmet, B. (2019). Classifying the reported ability in clinical mobility descriptions. *Proceedings of the 18th BioNLP Workshop and Shared Task*, 1–10. https://doi.org/10.18653/v1/W19-5001

Onal, K. D., Zhang, Y., Altingovde, I. S., Rahman, M. M., Karagoz, P., Braylan, A., … Lease, M. (2018). Neural information retrieval: at the end of the early years. *Information Retrieval Journal*, *21*(2), 111–182. https://doi.org/10.1007/s10791-017-9321-y

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Shao, Y., Mohanty, A. F., Ahmed, A., Weir, C. R., Bray, B. E., Shah, R. U., … Zeng-Treitler, Q. (2016). Identification and Use of Frailty Indicators from Text to Examine Associations with Clinical Outcomes Among Patients with Heart Failure. *AMIA Annual Symposium Proceedings*, 1110–1118. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/28269908

Skube, S. J., Lindemann, E. A., Arsoniadis, E. G., Akre, M., Wick, E. C., & Melton, G. B. (2018). Characterizing Functional Health Status of Surgical Patients in Clinical Notes. *AMIA Joint Summits on Translational Science Proceedings 2018*, 379–388. American Medical Informatics Association.

Sundar, V., Daumen, M. E., Conley, D. J., & Stone, J. H. (2008). The use of ICF codes for

information retrieval in rehabilitation research: An empirical study. *Disability and Rehabilitation*, *30*(12–13), 955–962. https://doi.org/10.1080/09638280701800285

Thieu, T., Camacho, J., Ho, P.-S., Porcino, J., Ding, M., Nelson, L., … Lai, A. M. (2017). Inductive identification of functional status information and establishing a gold standard corpus: A case study on the Mobility domain. *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2300–2302. https://doi.org/10.1109/BIBM.2017.8218042

US Social Security Administration. (2008). Disability Evaluation Under Social Security. Retrieved August 29, 2019, from 64-039 website: https://www.ssa.gov/disability/professionals/bluebook/general-info.htm

US Social Security Administration. (2014). Consultative Examinations: A Guide for Health Professionals. Retrieved August 29, 2019, from https://www.ssa.gov/disability/professionals/greenbook/index.htm

World Health Organization. (2001). *International Classification of Functioning, Disability and Health: ICF*. Geneva: World Health Organization.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V, Norouzi, M., Macherey, W., … Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv Preprint ArXiv:1609.08144*. Retrieved from http://arxiv.org/abs/1609.08144