Classifying the reported ability in clinical mobility descriptions

Denis Newman-Griffis[1,2]*, **Ayah Zirikly**[1]*, **Guy Divita**[1]*, **Bart Desmet**[1]
[1]Rehabilitation Medicine Department, Clinical Center, National Institutes of Health, Bethesda, MD
[2]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH
{denis.griffis, ayah.zirikly, guy.divita, bart.desmet}@nih.gov

Abstract

Assessing how individuals perform different activities is key information for modeling health states of individuals and populations. Descriptions of activity performance in clinical free text are complex, including syntactic negation and similarities to textual entailment tasks. We explore a variety of methods for the novel task of classifying four types of assertions about activity performance: *Able*, *Unable*, *Unclear*, and *None* (no information). We find that ensembling an SVM trained with lexical features and a CNN achieves 77.9% macro F1 score on our task, and yields nearly 80% recall on the rare *Unclear* and *Unable* samples. Finally, we highlight several challenges in classifying performance assertions, including capturing information about sources of assistance, incorporating syntactic structure and negation scope, and handling new modalities at test time. Our findings establish a strong baseline for this novel task, and identify intriguing areas for further research.

1 Introduction

Information on how individuals perform activities and participate in social roles informs conceptualizations of quality of life, disability, and social well-being. Importantly, activity performance and role participation are highly dependent on the environment in which they occur; for example, one individual may be able to walk around an office without issue, but experience severe difficulty walking along mountain paths. Thus, determining what level of performance an individual can achieve for activities in different environments is critical for identifying ability to meet work requirements, and designing public policy to support the participation of all people.

However, the interaction between individuals and environments makes modeling performance information a complex task. Assessments of activity performance within clinical healthcare settings are typically recorded in free text (Bogardus et al., 2004; Nicosia et al., 2019), and exhibit high flexibility in structure. Syntactic negation can be present, but is not necessarily indicative of inability to perform an action; for example, `Patient can walk with rolling walker` and `Patient cannot walk without rolling walker` are both likely to be used to assert the ability of the patient to walk with the use of an assistive device. Information about performance may also be given without a clear assertion, as in `the cane makes it difficult to walk`. Thus, extraction of performance information must not only distinguish between positive and negative assertions, but also those which cannot be clearly evaluated.

To the best of our knowledge, this is the first work to explore assertions of activity performance in health data. We explore a variety of methods for classifying assertion types, including rule-based approaches, statistical methods using common text features, and convolutional neural networks. We find that machine learning approaches set a strong baseline for discriminating between four assertion types, including rare negative assertions. While this work focuses on a relatively constrained and homogeneous corpus, error analysis suggests several broader directions for future research on classifying performance assertions.

2 Related Work

Though this is the first work focusing on the polarity of activity performance, three areas of prior work are particularly relevant to this research.

The first is concerned with applying NLP techniques and linguistic annotation to information about whole-person function, particularly activity

*These authors contributed equally to this work.

performance. Harris et al. (2003) experimented with term extraction for the purpose of terminology discovery to support information retrieval relating to functioning, disability and health, using linguistic, n-gram and hybrid techniques. Bales et al. (2005) and Kukafka et al. (2006) modified and applied the MedLEE NLP Extraction tool to code Rehabilitation Discharge Summaries using ICF (World Health Organization, 2001) encodings. Kuang et al. (2015) studied UMLS term coverage of functional status terms found in VA clinical notes and in social media sources, reporting that there is a need to extend existing terminologies to cover this area. Finally, Thieu et al. (2017) reported on an effort to build an annotated corpus of Physical Therapy (PT) notes from the Clinical Center of the National Institutes of Health (NIH) with functional status information. This corpus was also used for an investigation into using named entity recognition (NER) techniques to extract information about patient mobility (Newman-Griffis and Zirikly, 2018).

The second area is research on negation. Negation detection is a well-researched area (Morante and Sporleder, 2012), and both negation and uncertainty have historically been studied in the clinical NLP context (Mowery et al., 2012; Peng et al., 2018). Previous work studied the use of incorporating dependency parsers to help in identifying the scope (Sohn et al., 2012; Mehrabi et al., 2015). Recent work in this area involves the use of neural network models, where Long Short-Term Memory (LSTM), or variations of it, yielded competitive results on negation (cues and scope) detection (Taylor and Harabagiu, 2018).

One highly-related work to ours is Wu et al. (2014), which investigates detection of binary semantic negation status (i.e., the presence or absence of a finding, as opposed to syntactic negation) for clinical findings in EHR text. However, as Action Polarity is defined in terms of the interaction between an individual and a specific environment, it adds a layer of complexity to non-interactive physiological observations. Gkotsis et al. (2016) investigate using parsing-based scoping limitations for negation detection in complex clinical statements, though their focus is specifically on mentions of suicide.

Finally, classifying the assertion status of activity performance descriptions bears similarities to the problem of recognizing textual entailment (RTE) (Dagan et al., 2006; Marelli et al., 2014). RTE asks whether a given premise entails a specific hypothesis, and has historically been pursued in the general domain, though, recent efforts have developed datasets in biomedical literature (Ben Abacha et al., 2015; Ben Abacha and Demner-Fushman, 2016) and in clinical text (Romanov and Shivade, 2018). Our task, by asking whether a given description entails ability to perform an action in the an environment, is more constrained than RTE, but poses a related research challenge.

3 Data

We use an extended version of the dataset initially described by Thieu et al. (2017), consisting of 400 English-language Physical Therapy initial assessment and reassessment notes from the Rehabilitation Medicine Department of the NIH Clinical Center. These text documents have been annotated to identify descriptions and assessments of mobility status, typically including one or more specific Actions; for example, `Pt walked 300' with rolling walker` (Action underlined).

Each Action annotation was assigned one of four Polarity values, indicating what (if any) information the containing mobility description provides about the subject's ability to perform the given Action in the context of any described environmental factors.[1] The Polarity labels are defined in the following paragraphs.

Able The subject is able to complete the activity in the environment described. For example, `She states she can walk 20 minutes before tiring`; in the case of `now requires assistance of one person with transfers`, it is unknown whether the patient can perform the action independently, but they are able to do so with the assistance described.

Unable The subject is not able to complete the activity in the environment described; for example, `He is unable to walk`. More specific information may also be included, as in `Pt is now unable to walk more than 50 feet`.

Unclear Some information is provided about the subject's ability to perform the action, but not

[1] It is important to note that the Polarity label is dependent on the environmental factors described. For example, an individual may be able to walk a certain distance using an assistive device such as a rolling walker, but unable to walk that same distance independently.

Label	Train	Test	Total
Able	1,536	446	1,982
Unable	54	23	77
Unclear	158	48	206
None	1,784	478	2,262
Total	3,532	995	4,527

Table 1: Number of samples with each Polarity label in train and test data.

enough to make a definitive positive or negative judgment. For example, in `The cane makes it difficult to walk`, it is undetermined whether the subject can or cannot walk. This label also includes some cases of negated environmental factors; for example, `unable to propel wheelchair independently`.

None No direct information about ability to perform the action is provided. Common examples of this label refer to a scale that is either unavailable or distant in the document, as in `Ambulation: 1`. Other cases refer to a specific aspect of performing an action, without evaluation, as in `tendency during gait to quickly extend the leg from swing to stance`.

We randomly split the 400 documents into 320 training records and 80 testing records, stratified by distribution of Polarity labels. Table 1 provides frequencies of each label in these splits.

4 Methods

We investigate a variety of methods to classify the Polarity values of Action annotations. Rule-based methods have been used to great effect in clinical information extraction (Kang et al., 2013; Chapman et al., 2007), and form an important baseline for our task. We also make use of several common machine learning methods, such as support vector machines and k-nearest neighbors, along with more recent neural models such as convolutional neural networks (CNN). Finally, we experiment with ensembled combinations of our best-performing models. These approaches are described in the following subsections.

4.1 Rule-based

A UIMA (Ferrucci and Lally, 2004) based pipeline was constructed to identify action polarity from components of v3NLP-Framework (Divita et al., 2016). Leveraging the relationship of our task to detecting contextual attributes such as negation, the conTEXT (Chapman et al., 2007) algorithm embedded in the v3NLP-Framework was augmented with a few additional entries including "able" and "independent" as asserted evidence and "unable" as negative evidence.

The conTEXT algorithm relies on a lexicon of evidence and accompanying clues to indicate when evidence found to the right or left of a relevant entity within a bounded window should be applied. We used the sentence containing an Action mention as the bounds of its context window. An *Action Polarity* UIMA annotator was built to assign Polarity, given an Action annotation. This annotator is downstream from the conTEXT annotator that assigned negation, assertion, conditional, hypothetical, historical, and subject attributes to named entities. Within conTEXT-processed entities, we assigned *Unable* polarities to actions that had previously been attributed with negative and assigned *Able* polarities that had previously been assigned only asserted attributes. Actions that were tagged as conditional or hypothetical were not assigned a Polarity.

The v3NLP-Framework pipeline includes document decomposition annotators to identify sections, section names, sentences, slots and values, questions and their answers, and to a lesser extent checkboxes (Divita et al., 2014). Action mentions in clinical text occur within the boundaries of each of these elements. ConTEXT addresses action mentions within prose, but is not relevant for action mentions found in the semi-structured constructs. The Action Polarity annotator was thus augmented with additional rules to aid in polarity assignment based on where the mention was found. The most relevant rules are as follows:

- Action mentions that are in the slot part of a slot:value construct get their polarity assignment from positive or negative evidence in the value part of the construct. Table 2 provides guidelines to assigning polarity from slot:value and question and answer constructs.

- Action mentions that are within Goals or Education sections do not get a polarity. The section name is known for each named entity. For the time being, section names with "plan," "goals," "education," "intervention" and "recommendations" qualify. These are

Slot criteria	Value criteria	Assigned Polarity	Example
Asserted Action	Asserted Evidence	Able	`Transfers: Independent`
Asserted Action	Negated Evidence	Unable	`Transfers: Unable`
Negated Action	Negated Evidence	Able	`Difficulty Walking: No`
Negated Action	Asserted Evidence	Unable	`Unable to Walk: yes`
Asserted Action	Numbers	Unclear	`Transfers: 4`
Asserted Action	No context evidence	Unclear	`Sit to stand: minimal assist`
Asserted Action	No value	None	`Stand to sit:`
Multiple Actions	Doesn't matter	None	`Difficulty with chores, shopping, driving: Yes`

Table 2: Table of slot:value rules for Action Polarity

considered to be hypothetical constructs. The exception to this is if a goal is noted to have been met, it gets an *Able* Polarity.

- Action mentions within only the value part of the slot:value construct were handled the same way as Action mentions within prose.

4.2 Machine learning models

We evaluated the following common machine learning-based classification methods for our Polarity labeling task:[2]

- Random forest (RF), using 100 estimators;
- Naïve Bayes (NB), using Gaussian estimators;
- k-nearest neighbors (kNN), using k=5 with Euclidean distance;
- Support vector machine (SVM), with linear kernel;
- Deep neural network (DNN), using a 100-dimensional hidden layer followed by a 10-dimensional hidden layer.[3]

For a given Action mention a contained in a Mobility description m, we explored using both bag of binary unigram features[4] and word embedding features as model input. For both kinds of features, we experimented with using the context words in $m - a$ (i.e., all words in m except for the Action mention itself) only, and including the text of the Action mention a. Word embedding features were calculated by averaging the embeddings of all words used (either context alone or averaging context words and Action mention words together); we used pretrained FastText (Bojanowski et al., 2017) embeddings from Wikipedia and newswire, including subword information.[5] Where both unigram and embedding features are used, they are concatenated as a single feature vector.

Features	NB	RF	kNN	SVM	DNN
Unigrams	41.3	**77.3**	**67.0**	78.6	79.8
+Action	42.1	73.7	56.8	80.9	78.0
+Embeddings	41.6	64.3	66.3	78.8	**80.9**
+Both	**43.0**	65.1	65.2	**81.7**	79.6

Table 3: Macro F1 over Polarity classes in 5-fold cross validation feature selection experiments. All experiments start with binary unigram features using context words alone, and add Action words, embedding features from context words, or both (i.e., unigrams and embedding features from context and Action words combined). The best performing model configurations are marked in bold.

4.2.1 Feature selection

In order to identify the best combination of features for the task, we performed five-fold cross validation experiments on the training data. As shown in Table 3, we found that three model configurations achieved statistically equivalent macro F1 in cross validation ($p \geq 0.001$ with bootstrap permutation test, $R = 10000$ (Berg-Kirkpatrick et al., 2012)).[6] These are RF with unigram features (78.5% F1), the 2-layer DNN with unigram and embedding features from context only (80.9%), and SVM with all features, i.e. unigrams and embeddings with both the mobility description and Action mention texts (81.7%).[7]

Given the class imbalance in our dataset,

[2]We used the implementations of each method in Scikit-Learn (Pedregosa et al., 2011).

[3]We experimented with $d \in 10, 100$, and number of layers $\in 1, 2, 3$.

[4]Binary unigram features consistently matched or outperformed unigram counts in our experiments.

[5]`https://fasttext.cc/docs/en/english-vectors.html`

[6]We use significance threshold $p = 0.001$ throughout this paper, as a conservative Bonferroni correction for multiple testing. To have sufficient resolution to those low threshold, we use 10,000 replicates in bootstrapping.

[7]Complete results tables will be made available online.

Model	Able	Unable	Unclear	None	Macro
NB (All)	68.2	15.1	25.6	62.9	43.0
RF (Uni)	84.5	68.1	69.9	86.7	77.3
KNN (Uni)	73.5	53.3	62.6	78.5	67.0
SVM (All)	**86.3**	76.2	**76.4**	**87.8**	**81.7**
DNN (Uni+Emb)	85.0	**76.8**	74.3	87.5	80.9

Table 4: Five-fold cross validation results (F1) by class with best configurations of learned baselines. *All* indicates using unigrams, embeddings, and Action mention features; *Uni* indicates using unigram features from context words only, and *Uni+Emb* indicates both unigram and embedding features from context words. The best result in each column is marked in bold.

we also analyzed per-class performance of each model. Interestingly, as Table 4 illustrates, we found that all models except Naïve Bayes were surprisingly robust to this imbalance, with both SVM and DNN achieving over 76% F1 on the smallest class (*Unable*). Across all four classes, the SVM and the 2-layer DNN yield statistically equivalent performance ($p \geq 0.001$); we therefore use absolute macro F1 to choose SVM as the best baseline model for comparing across approaches.

4.2.2 CNN model

We adopt the Convolutional Neural Network (CNN) architecture introduced in Kim (2014). In our architecture, shown in Figure 1, we combine word embeddings with character embeddings, to reduce the impact of out-of-vocabulary rate as opposed to using words alone. Additionally, character-level CNNs have been shown to improve the results of text classification (Zhang et al., 2015), but the improvement is more evident with larger data sizes.

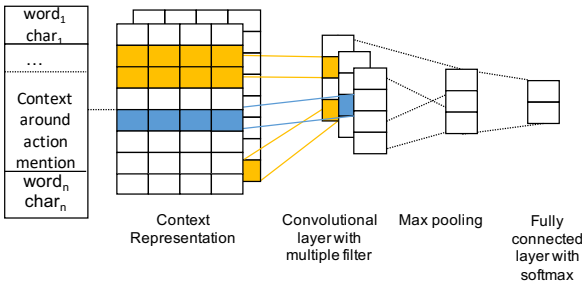


Figure 1: CNN architecture for Polarity classification.

Although our task is close to negation detection, it differs in that we do not need to detect the span of the Action: we take as inputs the Action mention and its parent mobility mention (a self-contained text span that can be considered a sentence). Unlike sequence tagging problems, where Long-Short Term Memory (LSTM) architectures would be a good fit (Fancellu et al., 2016), we treat the problem as a text classification task.

Embeddings	Able	Unable	Unclear	None	Macro
prev_all	82.3	48.7	31.8	86.9	62.4
next_all	79.3	32.3	53.5	82.7	64.9
full_all	**87.6**	**63.4**	65.0	**89.4**	**76.4**
full_char	66.0	45.7	**72.9**	78.7	65.8
full_word	86.1	42.4	60.3	88.0	69.2

Table 5: CNN performance using different inputs.

We experiment with character and word embeddings of the following inputs:

- previous context (*prev*): the set of words preceding and including the action mention.
- *next* context: the set of words following and including the action mention.
- *full* context: the union of $prev$ and $next$.

We also compare the impact of using character (*full_char*) or word (*full_word*) embeddings only as opposed to combining both (**_all*), as shown in Table 5. We note that relying on part of the context significantly drops the *Unable* performance. However, as expected, *prev* outperforms *next*, given that the words preceding the Action mention carry most of the ability-related information. For the rare *Unable* class, character embeddings outperform word embeddings, with F1 72.9% on the test set; the highest across all systems.

Hyperparameters were optimized on a dev set (we used a 90/10 train/dev split), yielding a learning rate of 0.0001, dropout of 0.5, embeddings size 100, and Adam optimization (Kingma and Ba, 2014) with L2 regularization.

4.3 Ensemble models

Ensembling methods have been shown to improve performance in a variety of classification tasks (Buda et al., 2018), including in class-imbalanced tasks (Ju et al., 2018). In order to combine the strengths of each modeling approach, we therefore experimented with ensembling all three systems, using two ensembling strategies:

Majority voting Predictions from the single best configurations of the SVM and CNN models[8] were combined to make a single decision. When

[8] Adding rule-based predictions degraded performance in this case.

System	Able			Unable			Unclear			None			Macro F1
	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1	
Rule-based	58.3	71.3	64.2	20.3	52.2	29.3	8.8	12.5	10.3	80.2	54.2	64.7	42.1
SVM	83.4	86.8	85.1	62.1	**78.3**	**69.2**	63.0	70.8	66.7	90.0	84.3	87.0	77.0
CNN	86.0	89.2	**87.6**	**72.2**	56.5	63.4	**81.2**	54.2	65.0	89.0	**89.7**	89.4	76.4
All (DNN chooser)	**87.5**	86.3	86.9	56.7	73.9	64.2	66.7	70.8	**68.7**	90.3	89.5	**89.9**	77.4
SVM+CNN (Voting)	82.3	**90.8**	86.4	62.1	**78.3**	**69.2**	62.1	**75.0**	67.9	**94.5**	82.2	87.9	**77.9**

Table 6: Precision (Pr), Recall (Rec), and F1 for each model evaluated on the test set. Top rows are individual models, bottom rows are ensembled results. The best result in each column is marked in bold.

the systems agreed, that label was chosen as output; in the case of disagreement, we chose the predicted label that was *less* frequent in training data, in order to prefer the strengths of individual models on rare classes.

DNN chooser Predictions from all three systems (rule-based and the best pretrained SVM and CNN models)[9] were passed as inputs to a DNN with a single 10-unit hidden layer.[10] In order to compensate for the class imbalance in our dataset, which would lead to preferring the CNN due to its higher precision, we identified all training samples that the three models disagreed on and grouped them by label, and identified the smallest of these disagreement sets. We then sampled no more than twice this number of points from each disagreement set, yielding a training sample of 182 points.

Using this downsampled training set, we trained the DNN to predict which, if any, of the systems chose the correct answer. As multiple systems may have made the correct prediction, this is a multi-label classification task. At test time, the system with highest probability output from the DNN was chosen as the reference decision for the final classification.

We also experimented with three approaches to predict the final class directly: using a DNN with the predictions of each system as input, using an SVM with predictions as input, and adding rule-based and CNN predictions as additional features to the SVM with lexical features. All variants underperformed the chooser in cross validation experiments on training data, thus we omit them from our results.

5 Results

The test results of the systems we compared are given in Table 6. The ensembled systems achieve the best overall performance, with 77.4% macro F1 with the DNN chooser and 77.9% with majority voting. Due in large part to the class imbalance in the dataset, the SVM, CNN, and ensemble methods do not yield statistically significantly different results in most cases ($p > 0.001$), although the voting ensemble does produce significantly higher precision on *None* samples than other methods ($p \ll 0.001$).

While performance is considerably better on the more frequent *Able* and *None* classes, the learned systems achieve good results on *Unclear* and the very rare *Unable*. Figure 2 shows the confusion matrices for all systems. The most common confusions are with *Able* and *None*, with only a small number of false positives for *Unable* and *Unclear* and no confusion between the two in the machine learning approaches.

Comparing between individual systems, the CNN is best at making the important distinction between *Able* and *Unable*. It consistently achieves high precision across all classes, but suffers large drops in recall for the rare labels. The SVM model reverses this tradeoff, yielding high recall for *Unable* and *Unclear*, but much lower precision. The ensembled methods are able to strike a good middle ground, keeping the high recall of the SVM without sacrificing too much of the CNN's precision.

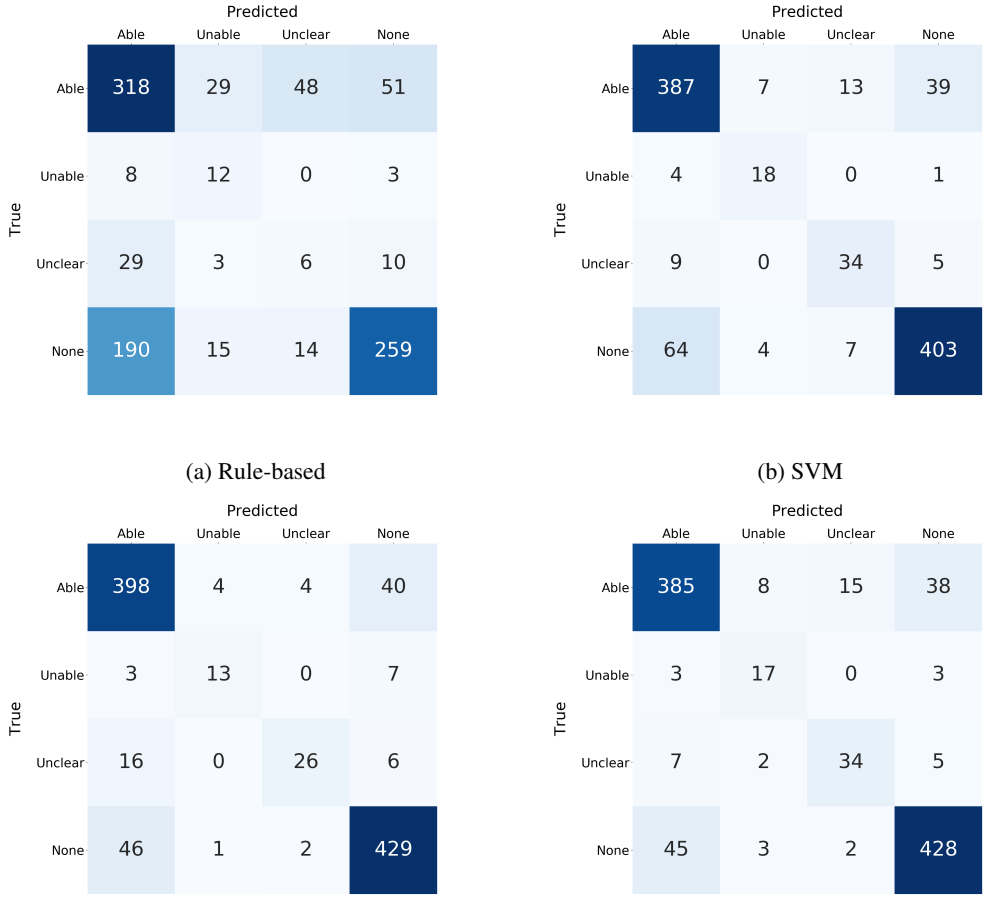
6 Discussion

As is evident from the results, correctly classifying the minority classes *Unable* and *Unclear* is not trivial. This is not only caused by the lack of data for training those classes, but in the case of *Unclear*, also by its semantic ambiguity – even for humans.

An important area of confusion is when actions are hypothetical, as is the case for plans, recommendations or feelings towards an action (e.g. `eager to walk`), which should all be tagged as *None*. Semantic problems can also arise around

[9] For the chooser, adding rule-based predictions consistently improved results over just SVM and CNN.

[10] Experiments with a 64-unit hidden layer, to cover all possible label combinations, yielded the same results in cross validation.



(a) Rule-based

(b) SVM

(c) CNN

(d) Ensemble (DNN chooser)

Figure 2: Confusion matrices for results on the test set.

the use of an assistive device. In the following synthetic example, the annotated polarity is *Able*: `she is unable to ambulate more than a few feet without support.` Without the mention of assistance, it would have been *Unable*. In future work, assistance mentions will be modeled explicitly to better capture this.

Overall, we obtain models that perform well across the board, where each approach has different strengths as illustrated in Figure 2. Out of the 955 test instances, the rule-based approach classifies 37 correctly that no other system got right. Likewise, SVM and CNN have 27 and 25 unique true positives, respectively. 46 instances get misclassified by all classifiers. The ensemble is able to pick up on 31 of the unique true positives from the machine learning systems, but consistently ignores valid suggestions from the rule-based approach. This suggests that different ensembling parameters should be considered to take better advantage of the rule-based system's strengths.

Below, we discuss system-specific observations in more detail.

6.1 Rule-based

The following failures were observed in the training and testing output:

Scoping negation The scope for assigning negation attribution was set to be within sentential boundaries. Ideally, the scope should be tighter at the major phrase level. However, v3NLP-Framework does not currently employ a dependency graph parser. Breaking on phrasal boundaries was not successful, primarily due to the inability to distinguish between list markers such as commas, coordinating conjunctions (and/or), and true scope limiting phrasal boundaries. Several false negatives were due to the incorrect *Unable* assignment because of negation scoping.

Identifying variants of slots and values accurately Negation and assertion assignment are dependent upon whether the action is within prose,

a slot or a value. A number of errors were due to multiple slot:value constructs within the same line making it difficult identifying the values, and/or nested constructs (i.e., the value of a slot:value construct was also a slot:value construct).

Nested sections A number of missed *None* errors were the result of mis-identifying what section the annotation was within, and picking up an inner section name. Several other issues arose from the use of spaces as delimiters between slots and values, as well as slots and values embedded within bulleted lists.

Pertinent negatives (Divita et al., 2014) A statement where the action mention had clear negative evidence really meant the patient could perform an action. For example, `no trouble walking`. An easy amelioration would be to gather constructs like "no trouble" and add them to the assertion evidence lexicon.

6.2 Machine learning

The machine learning systems are prone to failures in sentences that have multiple Action mentions, if their Polarity differs. This is because the systems do not take into account sentence structure. Similarly, sentence length seems to have a negative effect on performance, as it dilutes the information salient to the focus mention. In future work, we would limit the context information to exclude other mentions' contexts, add parse tree information relevant to the focus mention, or improve the neural network architecture to better model the sequential nature of the data.

The models would also benefit from better capturing semantic similarity. An example would be `Pt. is fearful to start walking again` (class: *None*), where the modality expressed by *fearful* might not have been learned from the training data. Additionally, lemmatization, stemming and character embeddings can blunt the impact of such unseen tokens, but using embeddings from large corpora would be more robust.

Finally, one potential limitation in our machine learning results is our use of pretrained embeddings from web text. As Newman-Griffis and Zirikly (2018) show, when only a small amount of text from the target domain is available, out-of-domain embeddings can roughly match performance with in-domain embedding features; however, developing or tuning more targeted word embeddings for use in this dataset is a useful area of future work.

6.3 Generalizability

It is important to note that the dataset used in this study was derived from one specialty – Physical Therapy – within a single institution – the NIH Clinical Center. Thus, the texts analyzed are likely to be more homogeneous than would be a broader dataset. Evaluating generalization of our findings to free text from other healthcare subdomains and other institutions, and describing ways in which performance assertions vary between these sources, is a valuable area of future work.

7 Conclusion

We have presented an evaluation of several approaches for the task of classifying whether a given description of an individual performing an activity indicates that they are able to perform it, unable, unclear, or insufficient information to determine. We found that machine learning approaches with lexical features perform surprisingly well on the task, including detecting the rarer labels of *Unable* and *Unclear*, and that an ensembled approach sets a strong baseline of 77.9% macro F1 for our dataset. In-depth analysis of system errors suggested several intriguing problems for future work. For instance, we intend to investigate hybrid models and test how information related to report formatting, section structure, slot info and assistive devices could improve the performance. To clarify the confusion of a patient's ability, we need models that can differentiate between factual and hypothetical statements (e.g. `Pt can run` vs. `Pt dislikes running`). Additionally, we would like to incorporate contextual representations such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018) into our models.

To our knowledge, this is the first work expanding on the problem of clinical negation detection to complex interactions between individuals and their environments. This work joins a growing body of research on application of NLP techniques to information about activity performance and role participation, and identifies several research challenges in adapting NLP methods to this new domain.

Acknowledgments

The authors would like to thank Pei-Shu Ho, Jonathan Camacho Maldonado, and Maryanne Sacco for discussions about error analysis, and our anonymous reviewers for their helpful comments. 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.

References

Michael Bales, Rita Kukafka, Ann Burkhardt, and Carol Friedman. 2005. Extending a medical language processing system to the functional status domain. In *AMIA Annual Symposium Proceedings*, volume 2005, page 888. American Medical Informatics Association.

Asma Ben Abacha and Dina Demner-Fushman. 2016. Recognizing question entailment for medical question answering. *AMIA Annual Symposium proceedings. AMIA Symposium*, 2016:310–318.

Asma Ben Abacha, Duy Dinh, and Yassine Mrabet. 2015. Semantic analysis and automatic corpus construction for entailment recognition in medical texts. In *Artificial Intelligence in Medicine*, pages 238–242, Cham. Springer International Publishing.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005. Association for Computational Linguistics.

Sidney T. Bogardus, Virginia Towle, Christianna S. Williams, Mayur M. Desai, and Sharon Inouye. 2004. What does the medical record reveal about functional status? *Journal of General Internal Medicine*, 16(11):728–736.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259.

Wendy W Chapman, David Chu, and John N Dowling. 2007. ConText: An algorithm for identifying contextual features from clinical text. In *Proceedings of the workshop on BioNLP 2007: biological, translational, and clinical language processing*, pages 81–88. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3944 LNAI:177–190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Guy Divita, Marjorie E Carter, Le-Thuy Tran, Doug Redd, Qing T Zeng, Scott Duvall, Matthew H Samore, and Adi V Gundlapalli. 2016. v3NLP framework: tools to build applications for extracting concepts from clinical text. *eGEMs*, 4(3).

Guy Divita, Shuying Shen, Marjorie Carter, Andrew Redd, Tyler Forbush, Miland N Palmer, Matthew H Samore, and Adi V Gundlapalli. 2014. Recognizing questions and answers in EMR templates using natural language processing. In *ICIMTH*, pages 149–152.

Federico Fancellu, Adam Lopez, and Bonnie Webber. 2016. Neural networks for negation scope detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 495–504.

David Ferrucci and Adam Lally. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

George Gkotsis, Sumithra Velupillai, Anika Oellrich, Harry Dean, Maria Liakata, and Rina Dutta. 2016. Don't let notes be misunderstood: A negation detection method for assessing risk of suicide in mental health records. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 95–105.

Marcelline R Harris, Guergana K Savova, Thomas M Johnson, and Christopher G Chute. 2003. A term extraction tool for expanding content in the domain of functioning, disability, and health: proof of concept. *Journal of biomedical informatics*, 36(4-5):250–259.

Cheng Ju, Aurélien Bibaut, and Mark van der Laan. 2018. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*, 45(15):2800–2818.

Ning Kang, Bharat Singh, Zubair Afzal, Erik M van Mulligen, and Jan A Kors. 2013. Using rule-based natural language processing to improve disease normalization in biomedical text. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):876–81.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jinqiu Kuang, April F Mohanty, VH Rashmi, Charlene R Weir, Bruce E Bray, and Qing Zeng-Treitler. 2015. Representation of functional status concepts from clinical documents and social media sources by standard terminologies. In *AMIA Annual Symposium Proceedings*, volume 2015, page 795. American Medical Informatics Association.

Rita Kukafka, Michael E Bales, Ann Burkhardt, and Carol Friedman. 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.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. SemEval-2014 Task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8, Dublin, Ireland. Association for Computational Linguistics.

Saeed Mehrabi, Anand Krishnan, Sunghwan Sohn, Alexandra M Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul Dexter, C Max Schmidt, Hongfang Liu, et al. 2015. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. *Journal of biomedical informatics*, 54:213–219.

Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Comput. Linguist.*, 38(2):223–260.

Danielle L. Mowery, Sumithra Velupillai, and Wendy W. Chapman. 2012. Medical diagnosis lost in translation: Analysis of uncertainty and negation expressions in english and swedish clinical texts. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, BioNLP '12, pages 56–64, Stroudsburg, PA, USA. Association for Computational Linguistics.

Denis Newman-Griffis and Ayah Zirikly. 2018. Embedding transfer for low-resource medical named entity recognition: A case study on patient mobility. In *Proceedings of the BioNLP 2018 workshop*, pages 1–11, Melbourne, Australia. Association for Computational Linguistics.

Francesca M Nicosia, Malena J Spar, Michael A Steinman, Sei J Lee, and Rebecca T Brown. 2019. Making function part of the conversation: Clinician perspectives on measuring functional status in primary care. *Journal of the American Geriatrics Society*, 67(3):493–502.

F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2018. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2017:188–196.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Sunghwan Sohn, Stephen Wu, and Christopher G Chute. 2012. Dependency parser-based negation detection in clinical narratives. *AMIA Summits on Translational Science Proceedings*, 2012:1.

Stuart J Taylor and Sanda M Harabagiu. 2018. The role of a deep-learning method for negation detection in patient cohort identification from electroencephalography reports. In *AMIA Annual Symposium Proceedings*, volume 2018, page 1018. American Medical Informatics Association.

Thanh Thieu, Jonathan Camacho, Pei-Shu Ho, Julia Porcino, Min Ding, Lisa Nelson, Elizabeth Rasch, Chunxiao Zhou, Leighton Chan, Diane Brandt, Denis Newman-Griffis, Ao Yuan, and Albert M Lai. 2017. Inductive identification of functional status information and establishing a gold standard corpus: A case study on the mobility domain. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2300–2302. IEEE.

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

Stephen Wu, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and Cheryl Clark. 2014. Negation's not solved: Generalizability versus optimizability in clinical natural language processing. *PLoS ONE*, 9(11).

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.