Toward Cross-Domain Engagement Analysis in Medical Notes

Adam Faulkner [*]
Grammarly
NY, NY, USA
adam.faulkner@grammarly.com

Sara Rosenthal
IBM Research
Yorktown Heights, NY, USA
sjrosenthal@us.ibm.com

Abstract

We present a novel annotation task evaluating a patient's *engagement* with their health care regimen. The concept of engagement supplements the traditional concept of adherence with a focus on the patient's affect, lifestyle choices, and health goal status. We describe an engagement annotation task across two patient note domains: traditional clinical notes and a novel domain, care manager notes, where we find engagement to be more common. The annotation task resulted in a κ of .53, suggesting strong annotator intuitions regarding engagement-bearing language. In addition, we report the results of a series of preliminary engagement classification experiments using domain adaptation.

1 Introduction

The recent trend in medicine toward health promotion, rather than disease management, has forefronted the role of patient behavior and lifestyle choices in positive health outcomes. Social-cognitive theories of health-promotion (Maes and Karoly, 2005; Bandura, 2005) stress patient self-monitoring of life-style choices, goal adoption, and the enlistment of self-efficacy beliefs as health promotive. We call this cluster of behavioral characteristics patient *engagement*. Traditional strategies of patient follow-up have also been affected by this trend: healthcare providers increasingly employ "care managers" (CMs) to monitor patient well-being and adherence to physician-recommended changes in health behavior—i.e., engagement. In this paper, we present an annotation schema for (lack of) engagement in CM notes (CMNs) and generalize the schema to the related domain of electronic health records (EHRs). Our high-level research questions are:

(1) Is the concept of engagement sufficiently well-defined that annotators can recognize the concept across text domains with an acceptable level of agreement?

(2) Can the annotations produced in (1) be used to classify engagement-bearing language across text domains?

In section 3, we report the results of our exploration of (1), describing an annotation task involving $\sim$ 6500 CMN and EHR sentences that resulted in an average κ of .53. In sections 4 and 5 we address (2) and report the results of several classification experiments that ablate classes of features and use domain adaptation to adapt these features to the CM and EHR target domains.

2 Related Work

The notion of patient engagement explored here is inspired by the self-regulation paradigm of (Bandura, 2005; Leventhal et al., 2012; Mann et al., 2013), where a patient's successful completion of health-related goals is predicated on their ability to "self-regulate", i.e., to plan and execute actions that promote attaining those goals, and their ability to maintain a positive attitude toward self-care. We are also aligned with the more recent work of Higgins et al. (2017) whose definition of engagement includes a "desire and capability to actively choose to participate in care".

NLP approaches assessing doctor compliance include Hazelhurst et al. (2005) who evaluate notes for doctor compliance to tobacco cessation guidelines and Mishra et al. (2012) who assess ABCs protocol compliance in discharge summaries.

[*] work completed at IBM

Proceedings of the BioNLP 2018 workshop, pages 189–193
Melbourne, Australia, July 19, 2018. ©2018 Association for Computational Linguistics

Label	Description	Examples
Engagement with care	The patient is engaged in their well-being by describing/exhibiting healthy behavior, positive outlook, and social ties.	"Patient disappointed by lack of weight loss but is just beginning exercise regimen"; "Patient joined book club."
Engagement with CM	Adherence to a doctor or CM instruction or understanding of CM advice.	"Patient verbalized understanding"; "Patient confided that she has gaps in nitroglycerin use."
Lack of engagement with care	Lack of engagement by using language suggestive of non-adherence to guidelines, health-adverse behavior, lack of social ties, or negative impression of patient self-care.	"White female, disheveled appearance"; "Patient admits to 'sedentary' lifestyle."
Lack of engagement with CM	Non-adherence to a prescribed instruction or a negative response to interaction.	"Patient rude during call"; "Patient angrily refused further outreach."
CM Advice	CM advice or suggestion	"I suggested he watch his diet and increase exercise"
Other	Default label to be chosen when no other label fits.	"Patient has a history of atrial fibrillation on corticosteroids"; "Chest is clear with no crackles."

Table 1: Annotation labels with descriptions and anecdotal examples. We use the term CM to describe both the para-professionals interacting with patients in CM notes and the physicians in EHRs.

While there exists work dealing with sentiment in clinical notes, such as positive or negative affect (Ghassemi et al., 2015) and speculative language (Cruz Díaz et al., 2012), (lack of) engagement cannot be reduced to sentiment. Lack-of-engagement-bearing language, for example, can also contain positive sentiment, e.g., *patient is feeling better so she has stopped taking her medication*. We include sentiment in our feature set, as described in Section 4.

The most closely related work is Topaz et al. (2017) who developed a document-level discharge note classification model that identifies the adherence of a patient in the discharge note. Their annotation task differs from ours, however, as they focus only on lack of adherence, specifically, towards medication, diet, exercise, and medical appointments. We also distinguish the targets of both engagement and lack of engagement by allowing annotators to identify either the CM or the care itself as the target.

3 Annotation Task and Data

The majority of our data consists of CMNs generated by a care manager service located in Florida, USA. CMs typically contact patients via phone to inquire into the patient's status with respect to health goals and enter the resulting information into the structured sections of a reporting tool. In addition, CMs note their impressions of the patient in a note as unstructured text, which we use here. To expand the domain scope of the task, we included EHR notes from the i2b2 Heart Disease Risk Factors Challenge Data Set (Stubbs and Uzuner, 2015; Stubbs et al., 2015), which includes notes dealing with diabetic patients at risk for Coronary Artery Disease (CAD). All notes were annotated in the same manner regardless of source.

3.1 Annotation Guidelines

Table 1 includes descriptions of the annotation labels along with anecdotal examples of each label type (original sentences are excluded due to privacy constraints[1]). Annotators were allowed to choose more than one label for each sentence, or no label at all (considered *other*). Our schema captures three different label classes: *engagement*, *lack of engagement*, and *cm advice*. We included *cm advice* because it can provide an indication that the next sentence should be classified as (lack of) engagement. We initially explored "barrier" language (e.g. *patient could not get to his appt because he didn't have a car*) as this can be indicative of lack of engagement, however, we found it to be too rare to include in the annotation tasks.

3.2 Annotation Challenges

Our first challenge was encoding a distinction between engagement and the more familiar concept of patient "adherence" (Vermeire et al., 2001; Topaz et al., 2017) in the annotation guidelines. While engagement-bearing language can include adherence-bearing language (e.g., *is monitoring blood sugar*, *made follow-up appointment*), the reverse is often not the case: Engagement-bearing language can include mentions of social ties (e.g., *discusses struggles to lose weight with sister*) and positive or negative evaluations of health-related goals (e.g., *patient was irritable when asked about efforts to reduce smoking*), neither of which involve adherence per se. By annotating such examples as engagement-bearing, we capture "self-

[1] All examples provided throughout the paper are anectodal

efficacy beliefs," which theories of patient self-regulation (Bandura, 1998, 2005) have suggested are predictive of health goal attainment.

An additional distinction that emerged during the annotation process involved the target of the engagement-bearing language: Is the patient (not) engaged with the CM or with the care itself? This distinction is evident in sentences that display a lack of engagement with care but a level of engagement with the CM. For example, in the sentence *He appeared cheerful in our interactions and admitted that he has not been exercising daily*, the patient is confiding in their CM (engagement) that they are not pursuing their health goals (lack of engagement). By allowing annotators to annotate such sentences as both engaged with the CM but unengaged with care we were able to exclude sentences that contained internally inconsistent engagement-bearing language from our data.

Another challenge involved the frequent use of "canned language" in the CM data, or language that does not report the CM's interactions with the patient but is used to meet some reporting criterion recommended by the health-care provider. For example, *Patient is scheduled for follow up appointment in two weeks*, is a frequently occurring canned language. Thus, we excluded common canned language sentences from the data.

3.3 Data Statistics

After several initial pilot rounds inter-annotator agreement for our six annotators on a final pilot round of 200 sentences (100 from each source) ranged from .46 to .66 among the annotators with an overall average of .53 (using Cohen's κ), indicating moderate to substantial agreement (McHugh, 2012).

4011 CMN sentences were annotated, extracted from $\sim 10,000$ unique CMNs. In order to broaden the range of language in our data, 2561 EHR sentences were annotated, with an equal number of sentences drawn from the three patient cohorts included in the i2b2 data. For each EHR, we restricted our annotation effort to sections that were more likely to include engagement-bearing language, specifically, the *social history*, *family history*, *personal medical history*, and *history of the present illness* sections. Table 2 shows the label distribution of the annotated data relative to note source. Although we allowed the annotators to differentiate between engagement/lack of engagement with care or the CM, we ultimately conflated these two categories into one for our experiments.

Source	**Engage**	**No Engage**	**Advice**	**Other**
EHRs	114	56	15	2376
CMNs	395	172	140	3304
Total	509	228	155	5680

Table 2: Label distribution relative to note type for all annotated sentence data.

4 Method

Given the small size of our data we elected to use a feature-engineering-based approach along with a discriminative classification algorithm in our experiments. Our features can be divided into five categories: *lexico-syntactic*, *lexical-count*, *sentiment*, *medical*, and *embeddings*.

Lexico-syntactic. Standard NLP features for text-classification such as n-grams and part-of-speech (POS) tags, along with dependency tuples (De Marneffe and Manning, 2008) with either the governor or dependent generalized to its POS.

Lexical-count. Frequency-based features such as sentence length, min and max word length, and number of out of vocabulary words.

Sentiment. We ran two sentiment classifiers over the data (Socher et al., 2013; Hutto and Gilbert, 2014) and included the resulting tags as features. In addition, we developed "comply word" features by inducing a lexicon based on WordNet- (Fellbaum, 1998) and Unified Medical Language System (UMLS)-based[2] synonym expansion of seed words such as "take" and "decline."

Medical. Using the MetaMap[3] tool, we generated Concept-Unique Identifiers (CUIs) for any medical concepts in the sentence. We also included both the "preferred names" and semantic types returned by UMLS for each concept

Embeddings. We extracted term-term, CUI-CUI and term-CUI co-occurrences pairs from a large medical corpus and used `wordtovecf`[4] (Levy and Goldberg, 2014) to learn embeddings from this co-occurrence dataset. We generated the mean of the embeddings for all content-words and CUIs in the sentence as a feature.

5 Experiments

All experiments were performed using an SVM classifier with a linear kernel basis function and

[2] http://www.nlm.nih.gov/research/umls/

[3] http://metamap.nlm.nih.gov/

[4] http://bitbucket.org/yoavgo/word2vecf

	experiment	CM			EHR		
		eng	lack	other	eng	lack	other
ALL	n-grams	18.4	25.1	91.3	7.1	3.3	95.4
	+embeddings	18.0	24.9	91.3	7.2	3.3	95.4
	+lexico-synt	22.7	20.7	90.2	7.0	3.2	94.6
	+lexical counts	21.7	21.9	89.2	9.1	3.1	93.1
	+sentiment	19.8	22.9	88.7	9.1	6.1	92.8
	+medical	22.3	24.3	89.3	8.4	8.6	93.6
	all	24.3	21.7	89.1	9.5	6.2	93.1
CM	n-grams	27.4	26.3	91.2	9.2	12.7	93.7
	+embeddings	27.5	26.8	91.3	8.5	12.7	93.7
	+lexico-synt	27.9	25.2	88.7	7.4	9.8	91.7
	+lexical counts	27.4	27.5	87.5	8.4	12.1	89.6
	+sentiment	27.6	**29.4**	87.4	7.9	11.3	89.4
	+medical	28.9	28.6	87.8	10.6	10.4	90.2
	all	**29.6**	25.4	87.1	10.8	**13.0**	89.6
EHR	n-grams	9.4	0.0	92.1	0.0	0.0	96.3
	+embeddings	9.4	0.0	92.1	0.0	0.0	96.3
	+lexico-synt	12.7	1.1	91.9	5.5	6.5	96.4
	+lexical counts	12.0	3.2	91.9	5.6	6.3	96.3
	+sentiment	10.2	3.2	91.9	6.9	6.3	96.5
	+medical	15.6	9.6	91.7	**17.0**	6.1	96.3
	all	11.1	7.2	91.7	14.4	6.3	96.2

Table 3: F-score results of ablation experiments for Engagement (*eng*) and Lack of Engagement (*lack*). Row headers refer to training sets and column headers refer to test sets. The best F-score for each test set is shown in **bold.**

one-vs-rest multiclass classification strategy as implemented in `scikit-learn`.[5] To deal with the skew in class distribution we experimented with both over- and under-sampling but got our best performance by simply adjusting class weights to be inversely proportional to class frequencies. Given the relatively small size of our data we used 5-fold cross-validation throughout. We also conflated *cm advice* with *other* to boost performance. We show F-score results for all three classes, but our analysis will focus on (lack of) engagement since *other* is trivially high-performing due to the massive data skew.

In our first set of experiments we examined the impact of training and testing on EHRs and CMNs individually, as well as together, while ablating the feature classes described in section 4. As shown in Table 3, all feature classes seem to help the model, but sentiment helps more for predicting lack of engagement in the CMNs while medical features help more for predicting lack of engagement in the EHRs. These experiments show that a CMN-trained model can perform well on EHRs. The best result for lack of engagement occurs when training on CM notes, with an F-score of 13.0.

The results in Table 3 encouraged us to apply domain adaptation (DA) to improve the results of the smaller dataset (EHRs) while also taking advantage of the larger dataset (CMNs) by considering EHRs to be "in-domain" and CMNs to be "out-of-domain". As this is still preliminary work, we started with a simple, yet effective DA strategy: the feature representation transformation procedure described in Daumé III (2007). Table 4 shows the results using DA. In the EHRs, where there is less data, on average DA provided an improvement, particularly for lack of engagement.

experiment	CM			EHR		
	eng	lack	other	eng	lack	other
n-grams	21.0	25.7	91.6	9.2	3.3	95.8
+embeddings	21.1	25.5	91.6	10.4	3.3	95.8
+lexico-synt	23.0	22.9	89.1	9.2	3.2	93.4
+lexical count	20.2	24.2	89.1	9.9	**15.0**	91.5
+sentiment	23.0	**26.1**	88.8	7.8	2.7	93.0
+medical	**23.4**	22.9	88.8	9.9	5.1	93.2
all	22.3	23.2	89.0	**13.3**	12.8	91.4

Table 4: F-score results of DA experiments. The best F-score for each test set is shown in **bold.**

6 Conclusion

In this paper we presented an annotation schema that captures engagement in CMNs and EHRs. We described the challenges of developing an annotation schema for a subjective task and show that annotators achieved moderate to high agreement in our final task. We annotated 6,572 sentences for (lack of) engagement and show preliminary results of a classification experiment on our dataset using feature ablation and domain adaptation. Our results are promising, showing that both features and domain adaptation are useful. However, they remain preliminary due to the rarity of (lack of) engagement labels. In future work, we plan to explore transfer learning to increase the size of our data, which in turn will allow use to explore deep learning approaches to this task.

Acknowledgements

Deidentified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organized by Dr. Ozlem Uzuner, i2b2 and SUNY. Research is based on care manager data provided by Orlando Health, a Florida-based community of hospitals.

[5] `http://scikit-learn.org`

References

Albert Bandura. 1998. Health promotion from the perspective of social cognitive theory. *Psychology and health*, 13(4):623–649.

Albert Bandura. 2005. The primacy of self-regulation in health promotion. *Applied Psychology*, 54(2):245–254.

Noa P Cruz Díaz, Manuel J Maña López, Jacinto Mata Vázquez, and Victoria Pachón Álvarez. 2012. A machine-learning approach to negation and speculation detection in clinical texts. *Journal of the Association for Information Science and Technology*, 63(7):1398–1410.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. *ACL 2007*, page 256.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8. Association for Computational Linguistics.

Christiane Fellbaum. 1998. A semantic network of english verbs. *WordNet: An electronic lexical database*, 3:153–178.

Mohammad M Ghassemi, Roger G Mark, and Shamim Nemati. 2015. A visualization of evolving clinical sentiment using vector representations of clinical notes. In *Computing in Cardiology Conference (CinC), 2015*, pages 629–632. IEEE.

Brian Hazlehurst, H. Robert Frost, Dean F. Sittig, and Victor J. Stevens. 2005. Mediclass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. *Journal of the American Medical Informatics Association*, 12(5):517–529.

Tracy Higgins, Elaine Larson, and Rebecca Schnall. 2017. Unraveling the meaning of patient engagement: A concept analysis. *Patient Education and Counseling*, 100(1):30 – 36.

C.J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*.

Howard Leventhal, Ian Brissette, and Elaine A. Leventhal. 2012. The common-sense model of self-regulation of health and illness. In *The self-regulation of health and illness behaviour*, pages 56–79. Routledge.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 302–308.

Stan Maes and Paul Karoly. 2005. Self-regulation assessment and intervention in physical health and illness: A review. *Applied Psychology*, 54(2):267–299.

Traci Mann, Denise De Ridder, and Kentaro Fujita. 2013. Self-regulation of health behavior: social psychological approaches to goal setting and goal striving. *Health Psychology*, 32(5):487.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. In *Biochemia medica*.

Ninad K. Mishra, Roderick Y. Son, and James J. Arnzen. 2012. Towards automatic diabetes case detection and abcs protocol compliance assessment. *Clinical medicine & research*, pages cmr–2012.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Amber Stubbs, Christopher Kotfila, Hua Xu, and Azlem Uzuner. 2015. Identifying risk factors for heart disease over time: Overview of 2014 i2b2 uthealth shared task track 2. *Journal of Biomedical Informatics*, 58:S67 – S77.

Amber Stubbs and Özlem Uzuner. 2015. Annotating risk factors for heart disease in clinical narratives for diabetic patients. *Journal of Biomedical Informatics*, 58:S78 – S91. Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

Maxim Topaz, Kavita Radhakrishnan, Suzanne Blackley, Victor Lei, Kenneth Lai, and Li Zhou. 2017. Studying associations between heart failure self-management and rehospitalizations using natural language processing. *Western journal of nursing research*, 39(1):147–165.

Etienne Vermeire, Hilary Hearnshaw, Paul Van Royen, and Joke Denekens. 2001. Patient adherence to treatment: three decades of research. a comprehensive review. *Journal of clinical pharmacy and therapeutics*, 26(5):331–342.