Predicting Therapist Empathy in Motivational Interviews using Language Features Inspired by Psycholinguistic Norms

James Gibson[1]*, Nikolaos Malandrakis*[1]*, Francisco Romero*[1]*,*
David C. Atkins[2]*, Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA, USA
[2]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA
[1]sail.usc.edu, [2]datkins@u.washington.edu

Abstract

Therapist language plays a critical role in influencing the overall quality of psychotherapy. Notably, it is a major contributor to the perceived level of empathy expressed by therapists, a primary measure for judging their efficacy. We explore psycholinguistics inspired features for predicting therapist empathy. These features model language which conveys information about affective and cognitive processes, which is central to the therapist expressing understanding of the patient's perspective. We describe the dimensional features obtained based on psycholinguisitic norms, and their application to predicting empathy expressed in motivational interviewing sessions for addiction counseling. We compare these to standard lexical features (n-grams) and demonstrate that these features contain complementary information for predicting therapist empathy. The highest empathy prediction results achieved are 75.28% UAR and 0.6112 Spearman's correlation.

Index Terms: behavioral signal processing, psycholinguistic norms

1. Introduction

With the growing prevalence of psychological interventions, it is vital to have measures which rate the effectiveness of psychological care providers. One such quality indicator is the behavioral construct of empathy, described as: "the extent to which the therapist understands and/or makes an effort to grasp the clients perspective" [1]. Expression of empathy helps therapists to establish a rapport with their patients, which leads to more effective interventions and ultimately more desirable outcomes [2, 3, 4]. Researchers from the psychology domain are interested in a deeper understanding, as well as establishing and improving the measurement, of relevant behaviors such as empathy [5, 6].

Xiao et al., analyzed the language of therapist empathy using common statistical language modeling techniques based on word frequency [7]. Their study used a subset of the same data we are using for this work, which will be described in the subsequent section. They found n-gram language models, trained on appropriately matched data sets, resulted in a maximum correlation of 0.56 with annotator assigned empathy ratings. Lord et al., analyzed language synchrony and its relation to empathy [6]. They reported that language style synchrony, a measure that is defined as both therapist and patient using the same semantic category in adjacent turns, was more predictive of empathy than counts of reflective utterances.

In this work we focus on therapist language in motivational interviews, a psychotherapy technique aimed at motivating patients toward achieving changes in behavior [8]. In particular, we consider motivational interviews which aim to curb alcohol and drug abuse by the participants receiving therapy. We analyze the language use of the therapists using novel computational constructs inspired by psycholinguistic norms [9] and relate it to their perceived level of empathy.

Language norms are numerical ratings which reflect the similarity of a particular word to various categories. Psycholinguistic norms specifically target linguistic representations of psychological processes such as affect and perception. Affective norms, typically with respect to the dimensions of *arousal*, *valence*, and *dominance*, are prevalent in state-of-the-art sentiment analysis [10]. There are also many well established norms that extend beyond affect, which aim to capture words' relations to cognitive processes, such as the age at which they are acquired (*age of aquisition*), the ability to form a mental image of the word (*imageability*), and even their degree of femininity or masculinity (*gender ladenness*) [11, 12]. Norms are usually manually annotated or automatically generated at the word level, with norms for sentences or turns estimated using functions on the norms of contained words [13]. Such dimensions provide an abstraction from words to categories which is missing in more common linguistic features such as n-grams. We use two methodologies to compute psycholinguistic norm features. The first is counts of words appearing in various psychological categories using a well established software tool. Secondly, we present novel features inspired by norms from psycholinguistic literature [11, 12, 14]. We evaluate the efficacy of these representations for predicting empathy and compare their performance to standard lexical features.

2. Motivational Interviewing Data

For this work, we use a corpus of motivational interviews (MI) collected from six independent clinical studies. Five of these studies consist of real patients who received interventions relating to alcohol, marijuana, and other drug use [5]. Three of these studies focus on alcohol use by young adults (ARC, ESPSB, and ESP21). The remaining two focus on use of marijuana (iCHAMP) and other drugs (HMCBI). All of these studies were behaviorally coded by trained annotators using the Motivational Interviewing Treatment Integrity (MITI), comprised of session level codes, [1] and Motivational Interviewing Skill Code (MISC), comprised of utterance level codes, [15] coding manuals. A subset of these data were also manually transcribed. These studies include 148 MI sessions in total.

In addition to the five clinical studies which are comprised of only real patients is the Context Tailored Training (CTT) data, which includes motivational interviews of both real and standardized patients [16]. Standardized patients are actors who are used to evaluate therapists for training purposes. These data include 200 fully transcribed and behaviorally coded sessions, 76 of these are with real patients and 124 are standardized patients (there are three unique standardized patients who appear in multiple MI sessions). This brings the total number of sessions in the Motivational Interviewing Corpus that we consider here to 348. Of these, three were removed from this analysis due to incomplete transcription or coding.

One of the primary behaviors of interest in this data is *empathy* [1]. The therapist performing the motivational interview in each session was rated for their level of *empathy* at the global (full session) level, on a Likert scale from 1-7. The distribution of *empathy* scores, y, is shown in Figure 1. There are 176 unique therapists in the six corpora. On average each therapist appears in approximately two sessions and the maximum number of sessions that any therapist appears is 12 (from the HMCBI corpus).

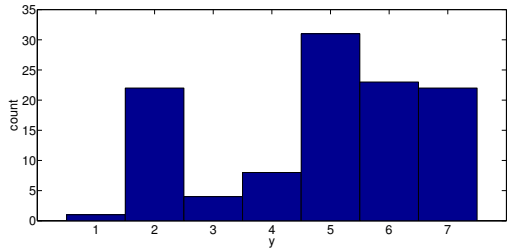

Figure 1: Empathy score distribution in MI Data

The corpus is separated into designated training and testing sets. They are split according to an approximate 70:30 training to testing ratio. Care was taken to ensure similar behavioral code distributions between training and testing splits, while placing all the sessions from each particular standardized patient into only one of the splits. This separation is because each standardized patient follows a particular story which they are acting out that will influence therapist language in their sessions.

3. Methodology

3.1. Predicting Therapist Empathy

In order to evaluate the efficacy of predicting therapist empathy from the MI data, we pose it first as a classification task then as a regression task. In addition to serving as a means for evaluating therapy efficacy, it is important to separate therapists who are displaying an acceptable to high level of empathy from those who do not to determine who should receive additional training.

We formulate this as a binary classification task by binarizing the session Likert scale ratings into classes of high and low empathy expression. The MITI coding manual states that a therapist much earn an *empathy* score of 5 or higher to be considered in adherence with motivational interviewing standards. For this reason, we consider all sessions with $y \geq 5$ as 'high empathy' and those with $y < 5$ as 'low empathy'. For all classification experiments we use linear Naive Bayes classifiers because they are common and have no parameters to tune. More complex classifiers could likely achieve higher accuracy but since we are chiefly concerned with understanding the relative contribution of each feature set we do not consider these options in this work.

On the other hand, for predicting the ranking of *empathy* scores we use the raw ratings, i.e., not the binarized labels. We estimate a generalized linear regression model using the features from the training data, then evaluate the model by using the Spearman correlation of its estimate of the *empathy* score, $\hat{y}$, with the annotator assigned score, y, on the testing data.

3.2. Description of Features

3.2.1. n-grams

In this work we consider unigram, bigram, and trigram word counts. Each higher order feature set includes the lower order sets (e.g., the bigram feature set includes counts of both single and pairs of words). Counting all the possible combinations of words and word sequences results in very large dimension feature vectors thus requiring feature selection/dimensionality reduction.

3.2.2. LIWC

The Linguistic Inquiry and Word Count tool (LIWC) computes the number of occurrences of words in particular categories of interest [17]. These categories include: parts of speech; psychological processes such as social, affective, cognitive, and perceptual constructs; personal concerns such as money, religion, and death; and spoken categories such as non-fluencies and fillers. LIWC compares the transcripts to a dictionary and counts the number of words belonging to each category. In this work only the psychological process dimensions are included, resulting in a 32 dimensional feature representation.

3.2.3. Psycholinguistic Norm Features

To compute features based upon psycholinguistic norms we use automatically generated word norms, created using an extension of the methodology in [13]. Following the algorithm, manually annotated norms are modeled based on their semantic similarity scores to highly frequent words, using a linear model, under the assumption that words of similar meaning (high semantic similarity) will have similar norms. Using this computational model we can create norms for new words (this methodology is described in greater detail in a companion submission to Interspeech 2015, [9]). The empirical psycholinguistic normative dimensions used in this research are presented in [11], [12], and [14]. We consider 13 psycholinguistic dimensions, including: Age of acquisition (aoa), Arousal (aro), Context Availability (conav), Concreteness (conc), Dominance (dom), Familiarity (fam), Gender Ladenness (gend), Imageability (imag), Meaningfulness-Colorado Norms (meanc), Meaningfulness-Paivio Norms (meanp), Pleasantness (pls), Pronouncability (pron), Valence (val). A short description of each norm is provided in Table 1.

For each of these dimensions, we also consider the part of speech (POS) of a particular token. So each raw feature represents the score of each psycholinguistic norm/POS pair, e.g., valence of plural nouns. The POS selection criterion can take any single value included in the Penn Treebank [18] plus some grouped tags for content words (all nouns, all verbs, all adjectives, all adverbs or all content words). Subsequently we compute functionals across each session to obtain a global feature representation. The functionals are: length (number of tokens), min, max, extremum (value furthest from zero), sum, average,

range, standard deviation, variance. Finally, for each feature we create a version that is normalized by the same value calculated over all unigrams, e.g., the maximum concreteness of past participles over the maximum concreteness of all unigrams. The full feature set has 10,998 feature dimensions (13 psycholinguistic dimensions x 47 POS filtering conditions x 9 functionals x 2 normalization schemes, normalized or unnormalized). We will subsequently refer to these as Psycholinguistic Norm Features (PNF).

Table 1: Description of psycholinguistic norms.

norm	description
aoa	expected age at which the word is acquired
aro	degree of excitement (versus calmness)
conav	number of contexts in which the word appears
conc	degree of concreteness (versus abstractness)
dom	degree of control over situation
fam	how commonly a word is experienced
gend	degree of femininity (versus masculinity)
imag	degree of ease in forming a mental image
mean-c,p	how associated a word is to other words
pls	degree to which pleasant feelings are associated
pron	degree of ease in pronouncing the word
val	degree of emotional positivity (versus negativity)

3.3. Feature Selection and Dimensionality Reduction

For the large feature sets (n-grams and PNF) we perform feature selection by ordering the features according to their information gain with the *empathy* rating. The information gain (or mutual information) is given by:

$$I(x;y) = \sum_x \sum_y p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right), \qquad (1)$$

where x is the feature and y is the *empathy* rating. In order to choose N_{IG}, the number of information gain ranked features retained from each set, we perform 2-fold cross validation on the training data.

For all feature sets we perform principal component analysis (PCA) using the sample covariance matrix given by:

$$\Sigma_{XX} = \frac{1}{n-1}(X - \bar{X})(X - \bar{X})^T, \qquad (2)$$

where n is the number of samples, and retain 99% of the variance, resulting in N_{PCA} feature dimensions. This is to reduce any redundancy between features, as well as to ensure the feature covariance matrix is non-singular. We show the original dimensionality of each feature set as well as the reduced dimensionality after feature selection and dimensionality reduction using information gain (N_{IG}) and PCA (N_{PCA}) in Table 2.

Table 2: Feature dimensionality.

features	orig.	N_{IG}	N_{PCA}
unigram	3,073	500	171
bigram	18,476	2,500	206
trigram	32,684	4,000	208
LIWC	32	NA	25
PNF	10,998	3,000	183

3.4. Feature Set Fusion

In order to determine whether the feature sets provide complementary information for the behavior prediction task we examine two methods of system fusions. The first is by concatenating the various feature sets, we will refer to this as early fusion. The second is by taking the mean of the posteriors of the classifiers trained on each feature set independently, we will refer to this as late fusion. For ranking we take the mean of the estimated scores from each regression model.

4. Experiments and Results

4.1. Prediction

We evaluate the efficacy of our various feature sets using two metrics: Unweighted Average Recall (UAR) to evaluate the success of binary classification and Spearman's Rank Correlation Coefficient (ρ) to evaluate the success of ranking the sessions. The UAR is an attractive metric for binary classification as the classes are heavily unbalanced (72.5% 'high empathy' vs. 27.5% 'low empathy' across the full corpus). We show prediction results in Table 3. This table also gives fusion results of the n-grams (n), LIWC (L), and PNF (P) feature sets.

The bigram feature set resulted in the highest UAR (without fusion) for prediction of *empathy*, while trigrams yielded the highest correlation. Features based on psychological norms result in competitive UAR to the n-gram feature sets. While the PNF and LIWC features result in lower Spearman's correlation, their fusion with n-gram features results in an improvement in both UAR and correlation. This demonstrates that the psycholinguistic feature sets (LIWC and PNF) carry complementary information for predicting *empathy*. It is also noteworthy that despite achieving lower performance than n-grams, the psycholinguistic groups of features contain the best individual indicators, and are critical for interpretability (presented in the next section).

Table 3: Prediction Unweighted Average Recall (%) and Spearman's correlation.

features	*empathy*	
	UAR	ρ
unigram	70.56	0.5399
bigram	**72.50**	0.5801
trigram	70.33	**0.6023**
LIWC	70.61	0.4428
PNF	71.33	0.4865
$\text{fusion}_{\text{n+L-early}}$	75.11	0.6016
$\text{fusion}_{\text{n+L-late}}$	70.72	0.5606
$\text{fusion}_{\text{n+P-early}}$	73.78	0.6027
$\text{fusion}_{\text{n+P-late}}$	71.44	0.6023
$\text{fusion}_{\text{L+P-early}}$	70.67	0.5049
$\text{fusion}_{\text{L+P-late}}$	69.20	0.5202
$\text{fusion}_{\text{all-early}}$	73.78	**0.6112**
$\text{fusion}_{\text{all-late}}$	**75.28**	0.5952

4.2. Feature Analysis

In Table 4, we show the features from each set which have the highest absolute Spearman's correlation with *empathy*. We present the top four features from each feature set, all were found to be significant at the 5% level (adjusted using the Bonferroni correction). These features lend some insight into the

types of expressions which are associated with *empathy*. For example, the top feature from all the n-gram sets includes the word 'sounds'. This word is often used in the phrase, 'it sounds like...', which is a common beginning to a therapist reflection. Reflections are behavioral acts which are highly correlated with *empathy* and is a behavioral code in the MISC manual. This is also captured in the *hear* and *perceptual* dimensions of the LIWC features, which is the top feature from that set. In addition to these perceptual categories, the other top LIWC features were affect related, *anxiety* and *affect*. This demonstrates that therapists discussing the feelings of the client, especially feelings of anxiety, play an important role in determining their level of *empathy*.

We examine the correlations between *empathy* and PNF features with respect to particular POS tags. The three parts of speech which were significantly correlated with *empathy* were JNRV (combination of adjectives, nouns, adverbs, and verbs; commonly referred to as content words), V (verbs, pertaining to actions), and N (nouns, pertaining to subject matter). The associations of *empathy* with *concreteness* and *imageability* are negative indicating that empathic therapists use more abstract language, which demonstrates therapists adhering to the client-centered aspect of motivational interviewing. This means that rather than telling a client to change their behavior, they attempt to influence them to develop and pursue their own goals for the behavior change. *Meaningfulness* is a measure of the ambiguity of a word. Its negative relation with *empathy* demonstrates the more unambiguous style of empathic therapists' language. Negative correlation of meaningfulness and context availability with *empathy* for verbs indicates discussion of actions specifically related to addiction. The *age of acquisition* of these words are positively associated with *empathy* as actions relating to addiction and substance abuse are typically of an adult nature. The negative relation of *empathy* to *pleasantness* is also a consequence of the nature of the actions being discussed. Addiction related verbs such as 'relapse' and 'abstain' exemplify these categories being unambiguous, of an adult nature, and associated with unpleasant feelings.

Table 4: Top features and their Spearman's correlation with *empathy*.

feature set	top features
unigram	sounds (0.32), ever (-0.31), severe (0.30), meds (0.28)
bigram	sounds like (0.33), to ten (0.33), severe risk (0.32), drug abuse (0.31)
trigram	it sounds like (0.31), in about a (0.30), abuse screening test (0.29), zero to ten (0.26)
LIWC	hear (0.36), perceptual (0.35), anxiety (0.30), affect (0.28)
PNF-JRNV	meanp (-0.37), conc (-0.31), imag (-0.28), conav (-0.23)
PNF-V	meanp (-0.43), conav (-0.42), aoa (0.40), pls (-0.35)
PNF-N	conc (-0.35), imag (-0.34), meanp (-0.23)

5. Conclusions and Future Work

We presented an approach for predicting therapist empathy using features inspired by psycholinguistic norms. We demonstrated that these features carry complementary information to standard lexical features (n-grams). While n-grams capture specific words and phrases that exemplify empathic behaviors these dimensions are more likely to capture therapist speaking style. Such features could be useful in measuring synchrony between therapist and patient, which is an important aspect of perceived level of therapist empathy [6].

We would like to build upon this work by evaluating these features using interaction models, i.e., determining similarities between therapist and patients as well as the dynamics of therapist behavior throughout the motivational interviews. Dynamic models will allow for modeling *empathic opportunities*, or patient utterances which invite empathic responses from the therapist. In addition to dynamic models, we are interested in estimating linguistic features directly from audio. This procedure will provide a fully automated system for evaluating important behaviors in psychotherapy. Such a system would be beneficial to the psychological community by drastically reducing the time necessary to assess these interactions. We are also interested in behavioral act tagging (BAT) to assign behavior tags to therapist utterances. These *behavioral acts* are specific behaviors which therapists employ in motivational interviews to facilitate the success of the intervention. They include, but are not necessarily limited, to behaviors from the MISC manual such as, *reflections*, *questions*, and *giving information*. An initial effort to detect *reflections* in motivational interviews is presented in [19].

6. References

[1] T. Moyers, T. Martin, J. Manuel, W. Miller, and D. Ernst, "The motivational interviewing treatment integrity (MITI) code: Version 2.0. university of new mexico, center on alcoholism," *Substance Abuse and Addictions (CASAA)*, vol. 2007, 2003.

[2] R. Elliott, A. C. Bohart, J. C. Watson, and L. S. Greenberg, "Empathy." *Psychotherapy*, vol. 48, no. 1, p. 43, 2011.

[3] J. Gaume, G. Gmel, M. Faouzi, and J.-B. Daeppen, "Counselor skill influences outcomes of brief motivational interventions," *Journal of Substance Abuse Treatment*, vol. 37, no. 2, pp. 151–159, 2009.

[4] J. McCambridge, M. Day, B. A. Thomas, and J. Strang, "Fidelity to motivational interviewing and subsequent cannabis cessation among adolescents," *Addictive Behaviors*, vol. 36, no. 7, pp. 749–754, 2011.

[5] D. C. Atkins, M. Steyvers, Z. E. Imel, and P. Smyth, "Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification," *Implementation Science*, vol. 9, no. 1, p. 49, 2014.

[6] S. P. Lord, E. Sheng, Z. E. Imel, J. Baer, and D. C. Atkins, "More than reflections: Empathy in motivational interviewing includes language style synchrony between therapist and client," *Behavior Therapy*, 2014.

[7] B. Xiao, P. G. Georgiou, and S. Narayanan, "Analyzing the language of therapist empathy in motivational interview based psychotherapy," in *Asia-Pacific Signal and Information Processing Association*, 2012, pp. 1–4.

[8] W. R. Miller and G. S. Rose, "Toward a theory of motivational interviewing." *American Psychologist*, vol. 64, no. 6, p. 527, 2009.

[9] N. Malandrakis and S. Narayanan, "Therapy language analysis using automatically generated psycholinguistic norms," in *companion submission to Interspeech*, 2015.

[10] S. Rosenthal, P. Nakov, A. Ritter, and V. Stoyanov, "Semeval-2014 task 9: Sentiment analysis in twitter," in *SemEval*, 2014.

[11] J. M. Clark and A. Paivio, "Extensions of the paivio, yuille, and madigan (1968) norms," *Behavior Research Methods, Instruments, and Computers*, vol. 36, no. 3, pp. 371–383, 2004.

[12] M. Wilson, "MRC psycholinguistic database: Machine-usable dictionary, version 2.00," *Behavior Research Methods, Instruments, and Computers*, vol. 20, no. 1, pp. 6–10, 1988.

[13] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, "Distributional semantic models for affective text analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2379–2392, 2013.

[14] M. M. Bradley and P. J. Lang, "Affective norms for english words (ANEW): Instruction manual and affective ratings," Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, Tech. Rep., 1999.

[15] W. R. Miller, T. B. Moyers, D. Ernst, and P. Amrhein, "Manual for the motivational interviewing skill code (MISC)," *Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*, 2003.

[16] J. S. Baer, E. A. Wells, D. B. Rosengren, B. Hartzler, B. Beadnell, and C. Dunn, "Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors," *Journal of Substance Abuse Treatment*, vol. 37, no. 2, pp. 191–202, 2009.

[17] J. W. Pennebaker, R. J. Booth, and M. E. Francis. (2007) Linguistic inquiry and word count: LIWC [computer software]. [Online]. Available: liwc.net

[18] M. Marcus, M. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of english: The penn treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, Jun. 1993.

[19] D. Can, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan, "A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features." in *Interspeech*, 2012.