Therapy Language Analysis using Automatically Generated Psycholinguistic Norms

Nikolaos Malandrakis, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory (SAIL), USC, Los Angeles, CA 90089, USA

malandra@usc.edu, shri@sipi.usc.edu

Abstract

Lexical norms, normative and usually numeric, ratings of word meaning are popular tools in research domains relating to human expression and perception of language, especially with regards to emotion. In this paper we are proposing an algorithm of psycholinguistic norm expansion capable of generating high quality norms representing aspects of language beyond emotion, including language concreteness and indicators of age and gender association. Starting from small manually annotated norm lexica, continuous norms for new words are estimated using semantic similarity and a simple linear model along eleven expression-related dimensions. The model is shown to achieve state of the art level performance of word norm estimation. To investigate the potential of these norms as analysis tools of more complex phenomena we use them to investigate the differences in therapist speech in sessions conducted by practitioners adhering to the psychoanalytic and client-centered schools of therapy.

Index Terms: natural language processing, computational linguistics, affective lexicon

1. Introduction

Psychotherapy has been increasingly popular in recent years as a subject of computational examinations, particularly with respect to the automatic evaluation of therapy quality. Studies have been conducted on discerning the differences between different methods/schools of therapy [1] and on modeling specific aspects of the therapist-patient interaction [2].

In this paper we focus on investigating the differences, in terms of therapist and patient language, between two popular schools of therapy: psychoanalytic psychotherapy (also known as psychoanalysis) and client-centered therapy. Psychoanalytic psychotherapy (henceforth referred to as *PP*), as defined by Sigmund Freud, targets the patient's subconscious mind. The therapist is meant to discover hidden or repressed memories, emotions and motivations over the course of therapy and use the gained insights to treat the patient. Client-centered therapy [3] (henceforth referred to as *CCT*), devised by Carl Rogers, is defined by the assumption that the patient has within himself the concepts of self-understanding and self-healing. The therapist's mission is to facilitate the self-healing process by providing an accepting and understanding environment. As part of that mission statement, CCT practitioners eschew the use of the word "patient" for "client", since the subject is viewed as a collaborator in the healing process. While the difference in definitions should translate into notable differences in therapist language, quantifying these differences is not simple. In this paper, we propose using language norms as a means to describe the language used and therefore any differences.

Language norms are (typically numeric) representations of the normative (expected) content of language, usually collected in a lexicon/thesaurus. The most commonly used variety are emotional norms, representing word polarity, valence, arousal or dominance. Manually annotated lexica [4, 5] have limited computational applications due to their small sizes (typically a few hundred up to a few thousand words), so machine learning methods are used to expand them and create automatically generated lexica [6]. Emotional norms have been utilized for a variety of tasks and remain at the core of cutting edge sentiment analysis systems [7]. Norms for aspects of language beyond emotion have existed for a long time [8, 9] and have been popular in behavioral studies as a way to select appropriate stimuli [10]. These include norms describing the degree of abstractness, the complexity of meaning and age or gender affinity of individual words, so there should be potential for a variety of uses. Very recently some of these norms have become the target for expansion and used in computational analyses [11, 12, 13] though they still are not particularly common.

In this paper we are proposing a method of expanding psycholinguistic norms using a variation of a method of emotional lexicon expansion that we had previously proposed [14]. The method is shown to achieve state-of-the-art prediction performance when applied to eleven different dimensions, most of which have never before been automatically generated. Then we use the expanded psycholinguistic norms to quantify aspects of the difference between client centered therapy and psychoanalysis.

The method used to expand word-level psycholinguistic norms is detailed in Section 2. The extra steps used to create norms per sentence, turn or other larger units are described in Section 3. The corpora and methodology used in the experiments are presented in Section 4 and Section 5. Results are presented and discussed in Section 6. Finally conclusions and future work are discussed in Section 7.

2. Expanding Word Psycholinguistic Norms

The method of generating word norms is an extension of the approach described in [14], which in turn is a generalization of the method presented in [15]. At its core is the assumption that semantic similarity can be used to estimate similarity in any other domain: words with similar meaning will have similar norm values. More specifically, we assume that the norm value $v(t_j)$ corresponding to term t_j can be estimated through a linear combination of the semantic similarities $s(t_j, w_i)$ between the term and known seed words w_i and the norms $v(w_i)$ using the equation

$$\hat{v}(t_j) = a_0 + \sum_i a_i \cdot s(t_j, w_i) \cdot v(w_i), \qquad (1)$$

where a_i are trainable weights corresponding to words w_i. Starting from a manually annotated lexicon containing K words, we can select $N < K$ words as seed words and create a system of K equations with $N + 1$ variables ($a_0 ... a_N$) that can be solved using Least Squares Estimation (LSE).

This method works very well, producing state-of-the-art results when estimating emotional norms [14]. Applying to a larger and more varied set of norms complicates one aspect of the process: seed word selection. The seed words w_i have to be selected separately for each norm, from the words available in the original manual annotations, a process that may be affected by the nature of the norm in some way. Also, since norm lexica from different sources contain different words, the number of similarity ratings required increases dramatically.

A solution can be reached via a simple modification to the method. Since the weights a_i are trainable, they will compensate for the absence of $v(w_i)$. For example, if a norm has a value of x and we eliminate it, the corresponding trained coefficient will increase in value by a factor of x. Therefore we eliminate all $v(w_i)$ and get the new equation:

$$\hat{v}(t_j) = a_0 + \sum_i a_i \cdot s(t_j, w_i), \tag{2}$$

which allows us to use *any* set of words, multi-word terms or concepts as seeds, independently of the norm we are estimating.

The similarity metric $s()$ used in (2) is a critical component of the process. Through the course of prior experiments [14] we found context based similarity to provide the best performance, a finding that holds for norms beyond emotion. The metric used is the cosine between context vectors with binary weights and a window size of one: the context vector for word w_i contains ones in all places corresponding to terms that occur right next to it (window size of one) at least once in some large corpus. For a detailed description of the experiments that lead to the selection of this similarity metric, as well as the performance impact on emotion norm estimation, see [14].

3. Norms for larger passages

Given the word norm model, we can generate norms for any word that occurs in any particular corpus we want to process. Still, creating norms for larger lexical units such as sentences or turns is not straightforward.

The dominant approach in emotion related literature, and our prior work, is extracting statistics that describe the word norm distribution. The target passage is tokenized and filtered, then a lexicon lookup is performed and words are replaced by their norms before statistics are estimated. The most common variant of that approach includes part-of-speech tagging, then removal of all words that are not content words (adjectives, nouns, verbs and adverbs), followed by extracting the average norm value across the content words. More complex variants can generate large numbers of features via combining multiple filtering criteria (such as multiple different part-of-speech selection criteria) with multiple statistics, including the minimum, maximum, slopes etc. Mapping these multiple statistics to passage norms can be achieved using machine learning.

For the purposes of this paper, the passage norm is estimated as the average norm across content words.

4. Corpora and Experimental Procedure

Following is a short description of the datasets used for the experiments presented in this paper.

4.1. Manually annotated norms

In order to generate psycholinguistic norms we need to start from some manual annotations. To that end we use eleven dimensions from three sources. More norm dimensions are available in these resources: the ones selected for expansion were picked based on their projected utility for our applications in behavioral informatics. Following is a list and short description for each dimension of interest.

From the Affective Norms for English Words (ANEW) [4] we use the norms for arousal, valence and dominance, the three dimensions of the widely used dimensional model of affect [16]. *Arousal* represents excitement, the degree of physical activation in preparation for action. *Dominance* is the degree of perceived control over one's circumstances. *Valence* is the continuous polarity (from very negative to very positive) of an emotion.

From the MRC Psycholinguistic database [9] we use the norms for concreteness, imagability, age of acquisition and familiarity. *Concreteness* is the degree to which something can be perceived using the five senses (from very abstract to very concrete). *Imagability* is the degree to which one may create a mental image of the word's subject. *Age of acquisition* indicates the expected age at which one acquires (can use correctly) the word. *Familiarity* is the degree of exposure to and knowledge of the word.

From the Paivio, Yuille and Madigan norms [8] we use the norms for pleasantness, pronouncability, context availability and gender ladenness. *Pleasantness* is very similar to valence and is a degree of how pleasant the feelings associated with a word are. *Pronouncability* signifies how easy a word is to pronounce. *Context availability* represents the number of different contexts in which a word may appear. *Gender Ladenness* represents the degree of perceived feminine or masculine association of a word (from very masculine to very feminine).

4.2. Raw text corpus

The norm generating equation (2) requires a semantic similarity estimate $s()$. In this paper semantic similarity is calculated as the cosine of context vectors calculated over a large corpus of raw text. The corpus was created by posing a query to the Yahoo! search engine for every word in the English version of the [17] spell-checker and collecting the top 500 result previews. Each preview is composed of a title and a sample of the content, each being a single sentence. Overall the collected corpus contains approximately 117 million sentences.

4.3. General psychotherapy corpus

The general psychotherapy corpus, maintained and updated by the "Alexander Street Press" (http://alexanderstreet.com/) and made available via library subscription, contains transcripts of therapy sessions. Over 1200 sessions in total are included, spanning a large variety of therapeutical approaches and experimental conditions. Beyond the therapist and client transcripts, the corpus includes metadata such as therapist and client demographics and client symptoms.

As noted in other studies [18], this corpus has some unfavorable characteristics. It is missing a lot of metadata, with most only available for a minority of the provided sessions, and it is unbalanced with respect to almost all variables since it was not specifically designed for analysis as proposed here. Despite that, the large size makes it a valuable resource.

For these experiments we used the therapist and client language, the therapist school of therapy (if it is CCT or PP) and the therapist's experience (under 10 years, 11-20 years). The subset of the original corpus that fits these criteria contains 312 therapy sessions.

Table 1: Word norm estimation performance. Cardinality of the dataset and Pearson correlation to the ground truth for 10-fold regression over manually annotated words.

Dimension	#Samples	Pearson
Arousal	1034	0.70
Dominance	1034	0.77
Valence	1034	0.88
Pleasantness	925	0.82
Concreteness	4295	0.87
Imagability	4829	0.85
Age of Acquisition	1904	0.86
Familiarity	4924	0.86
Pronouncability	925	0.72
Context Availability	925	0.78
Gender Ladenness	925	0.80

5. Experimental Procedure

To evaluate the performance of the word norm expansion algorithm we performed 10-fold cross-validation experiments for each norm dimension. In all cases we rescaled the original norm values to the range $[-1, 1]$. The seed words w_i of (2) were selected using word frequencies calculated over the corpus described in section 4.2. The top 10000 most frequent words with length longer that 3 characters were used as seeds for all experiments. Given each training set, we generated the $K \times 10000$ matrix of similarities to the seed words. Dimensionality reduction was performed by applying Principal Component Analysis (PCA) and keeping the first N components. These transformed similarities became the similarity terms $s()$ of (2) which in this case represent similarities to concepts (combinations of words), rather than individual words. For the experiments presented in this paper, the first 500 principal components were used for the MRC and ANEW sourced norms, whereas only the first 300 components were used for norms taken from the smaller Paivio-Yuille-Madigan resource.

To generate norms for any new words we trained models using all manually annotated samples as training samples. The resulting model for each norm dimension is composed of a coefficient vector containing all a_i and a PCA transformation matrix, required to map from the original semantic space of 10000 dimensions. For any new word we calculated the 10000 similarities to the seed words, transformed using the trained PCA matrices and plugged into the learned equations to get the corresponding norms.

The analysis of therapy transcripts started by part-of-speech tagging the patient and client utterances with Treetagger [19]. Non-content words were removed and the content word vocabulary was created. For each word in that vocabulary we generated all possible norms and calculated the norm averages *per session*. The result was a eleven norms *per session*, per interlocutor.

The target of analysis was the therapist language given the client language. To do that we estimated the therapist norms given the client norms, assuming a linear relationship, and kept the residuals (the errors). These estimation errors are attributed to the therapist and become the dependent variable of our analysis. The independent variables are: the school of therapy the therapist belongs to and the years of experience he or she has. Using more independent variables was desirable, but was not possible due to the unbalanced corpus that resulted in empty groups. The method of analysis was Welch's ANOVA, a variant of ANOVA for unequal variances.

6. Results

6.1. Word-level norm estimation

The performance of the word norm expansion algorithm is shown in Table 1. For each dimension the number of manually annotated words and the Pearson correlation to the ground truth are listed. Overall the model performs very well, with Pearson coefficients around 0.8 for almost all dimensions. Arousal from language is predictably harder to estimate and pronouncability is not strictly defined by semantics alone, so the relatively low performance of the model on these dimensions was expected.

Most of these dimensions have not been automatically expanded before, so comparisons are difficult. The 0.88 Pearson for valence is higher than the 0.87 cited in [14], the 0.87 Pearson for concreteness is higher than the 0.81 cited in [13] and the 0.85 imagability is higher than the 0.56 cited in [12]. Therefore this system performs better overall than related attempts in literature.

Table 2: Pearson correlation of Client and Therapist norms, for Client centered therapy, psychoanalytic psychology and Overall.

	CCT	PP	Overall
Arousal	0.15	0.08	0.18
Dominance	0.20	0.20	0.19
Valence	0.20	0.24	0.19
Pleasantness	0.15	0.22	0.18
Concreteness	**0.47**	**0.26**	0.32
Imagability	**0.43**	**0.17**	0.24
Age of Acquisition	**0.37**	**0.19**	0.29
Familiarity	0.30	0.33	0.29
Pronouncability	0.31	0.32	0.29
Context Availability	0.32	0.26	0.24
Gender Ladenness	0.35	0.32	0.30

6.2. Therapy transcript analysis

As a first step, we looked at the relation between the language norms of the therapist and client, by calculating the Pearson correlation between their per-session features, for each school of therapy and overall. The results are presented in Table 2. All correlation over 0.15 are significant at the $p < 0.05$ level and correlations over 0.25 are significant at the $p < 0.001$ level. As expected, there is significant correlation between therapist and client along almost all dimensions. The differences in correlation with the client between the two schools of therapy are not significant with the exception of concreteness, imagability and age of acquisition where the CCT therapists correlate better. The finding seems consistent with the school descriptions: the CCT therapists should follow and empower the client, while the PP therapists have specific goals that may conflict with the short-term goals of the client (what he or she wants to discuss).

To compare therapist language we perform statistical analysis on their norms, normalized by client norms using linear regression, with respect to the school of therapy (SoT) and the

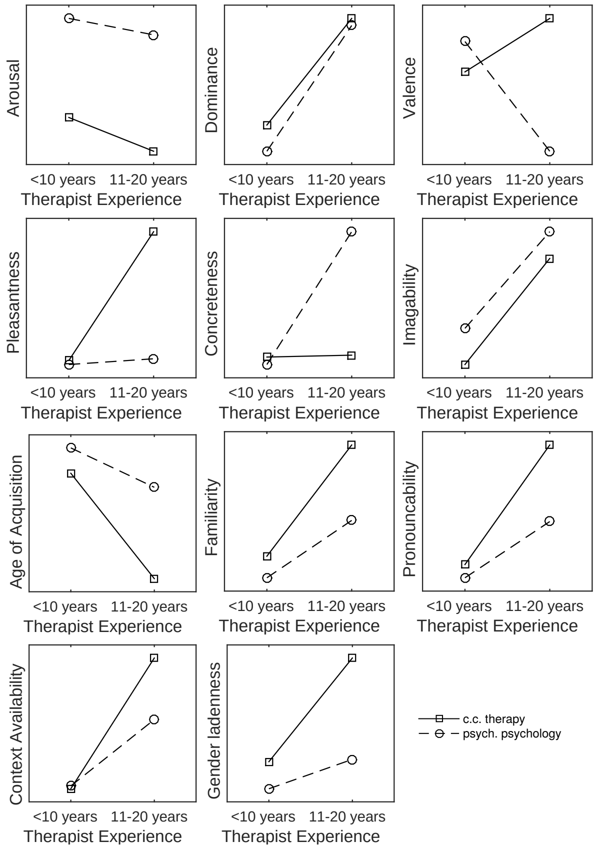

Figure 1: Sample means as functions of therapist experience for client-centered therapy and psychoanalytic psychology.

therapists' experience (TE). The results are presented in Table 3. Most of these differences have a strong interaction component that warrants further investigation before commenting on main effects. One exception is dominance, the difference of which is attributed to higher therapist experience, an indicator that a therapist with more experience gives a stronger impression of being in control.

To investigate the interaction of SoT and TE we created the factor plots shown in Fig. 1 and ran pair-wise statistical tests for all values of the interacting terms. This investigation reveals only one main effect, only one factor that differentiates between schools for all levels of experience: arousal, with PP transcripts showing higher values. Another way to interpret this would be that the CCT practitioners use language consistent with higher calmness levels.

Beyond arousal, all other effects only become significant at higher experience levels. The effect of increased experience on the two schools is shown in Table 4, that also includes the differences between schools at the higher experience level. Experience appears to have much more of an effect on CCT than PP, with CCT therapist language becoming more pleasant (higher pleasantness), simpler (lower age of acquisition, higher pronouncability) and more accessible (higher familiarity, higher context availability). A lot of these trends are also evident for PP practitioners, as seen in Fig. 1, but none are significant, apart from a significant increase in concreteness. The differences between the two schools at the high level of experience mirror the changes for CCT: the therapist language in CCT transcripts is simpler, more accessible and more pleasant than the language in PP, in addition to the main affect of appearing more relaxed.

The observed differences seem consistent with the CCT target of providing an "accepting and understanding" environment for the client.

Table 3: Factor p-values and direction of difference. Significant differences are denoted with $\uparrow$ or $\downarrow$ at the $p < 0.05$ level and $\Uparrow$ or $\Downarrow$ at the $p < 0.001$ level.

	p-value			direction	
	SoT	TE	SoT*TE	PP	TE$\uparrow$
Arousal	0.001	0.033	0.001	$\Uparrow$	$\uparrow$
Dominance	0.167	0.041	0.200		$\uparrow$
Valence	0.071	0.060	0.091		
Pleasantness	0.658	0.494	0.001		
Concreteness	0.003	0.002	0.018	$\uparrow$	$\uparrow$
Imagability	0.004	0.001	0.017	$\uparrow$	$\Uparrow$
Age of Acquisition	0.928	0.059	0.000		
Familiarity	0.274	0.002	0.000		$\uparrow$
Pronouncability	0.138	0.001	0.000		$\Uparrow$
Context Availability	0.017	0.000	0.000	$\uparrow$	$\Uparrow$
Gender Ladenness	0.613	0.470	0.003		

Table 4: The effect of increased experience. Significance denoted with $\uparrow$ or $\downarrow$ at the $p < 0.05$ level and $\Uparrow$ or $\Downarrow$ at the $p < 0.001$ level.

	CCT	PP	CCT vs PP
Arousal			$\Downarrow$
Dominance			
Valence			$\uparrow$
Pleasantness	$\Uparrow$		$\Uparrow$
Concreteness		$\uparrow$	
Imagability			
Age of Acquisition	$\Downarrow$		$\Downarrow$
Familiarity	$\Uparrow$		$\Uparrow$
Pronouncability	$\Uparrow$		$\Uparrow$
Context Availability	$\Uparrow$		$\uparrow$
Gender Ladenness			

7. Conclusions

We proposed and evaluated a method for creating psycholinguistic norms for words based on manually annotated lexica and semantic similarity. The method assumes a linear relation between semantics and all other aspects and is trained using LSE. The method achieved state-of-the-art results on word-level norm generation. The norms were used to analyze the differences between different schools of therapy and the findings were consistent with the theoretical definitions of the schools, with client-centered therapist speech appearing simpler, calmer and more pleasant. Future work will include the expansion of the norm model to more dimensions, which will enable a more detailed description of language. At the analysis level, moving to sentence or turn-level analysis could help provide further insights. Finally, future work will apply the approach presented here in conjunction with analysis of other related data (e.g., speech) to obtain a more comprehensive account of the mechanisms, quality and efficacy of psychotherapy within a behavioral informatics framework [20].

8. Acknowledgements

Work supported by NSF, NIH and DoD.

9. References

[1] H. Strupp, "A multidimensional comparison of therapist activity in analytic and client-centered therapy." *Journal of Consulting Psychology*, vol. 21, no. 4, pp. 301–308, 1957.

[2] D. Can, P. Georgiou, D. Atkins, and S. Narayanan, "A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features," in *Proc. Interspeech*, 2012, pp. 2254–2257.

[3] C. R. Rogers, *Client-centered Therapy*. Boston: Houghton Mifflin, 1951.

[4] M. Bradley and P. Lang, "Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Technical report C-1." The Center for Research in Psychophysiology, University of Florida, 1999.

[5] P. Stone, D. Dunphy, M. Smith, and D. Ogilvie, *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press, 1966.

[6] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proc. LREC*, 2006, pp. 417–422.

[7] S. Rosenthal, P. Nakov, A. Ritter, and V. Stoyanov, "SemEval-2014 Task 9: Sentiment analysis in Twitter," in *Proc. SemEval*, 2014.

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

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

[10] P. J. Schwanenflugel, C. Akin, and W.-M. Luh, "Context availability and the recall of abstract and concrete words," *Memory & Cognition*, vol. 20, no. 1, pp. 96–104, 1992.

[11] T. Liu, K. Cho, G. A. Broadwell, S. Shaikh, T. Strzalkowski, J. Lien, S. Taylor, L. Feldman, B. Yamrom, N. Webb, U. Boz, I. Cases, and C.-S. Lin, "Automatic expansion of the MRC psycholinguistic database imageability ratings," in *Proc. LREC*, 2014.

[12] M. Mohler, M. Tomlinson, D. Bracewell, and B. Rink, "Semi-supervised methods for expanding psycholinguistics norms by integrating distributional similarity with the structure of wordnet," in *Proc. LREC*, 2014.

[13] P. Turney, Y. Neuman, D. Assaf, and Y. Cohen, "Literal and metaphorical sense identification through concrete and abstract context," in *Proc. EMNLP*, 2011.

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

[15] P. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Transactions on Information Systems*, vol. 21, pp. 315–346, 2003.

[16] J. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, no. 3, pp. 273 – 294, 1977.

[17] "Gnu aspell," http://www.aspell.net.

[18] Z. Imel, M. Steyvers, and D. Atkins, "Computational psychotherapy research: Scaling up the evaluation of patient-provider interactions," *Psychotherapy*, vol. 52, no. 1, pp. 19–30, 2015.

[19] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proc. International Conference on New Methods in Language Processing*, vol. 12, 1994, pp. 44–49.

[20] S. S. Narayanan and P. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of IEEE*, vol. 101, no. 5, pp. 1203 – 1233, May 2013.