Sentence level estimation of psycholinguistic norms using joint multidimensional annotations

Anil Ramakrishna, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, CA
akramkr@usc.edu, shri@ee.usc.edu

Abstract
Psycholinguistic normatives represent various affective and mental constructs using numeric scores and are used in a variety of applications in natural language processing. They are commonly used at the sentence level, the scores of which are estimated by extrapolating word level scores using simple aggregation strategies, which may not always be optimal. In this work, we present a novel approach to estimate the psycholinguistic norms at sentence level. We apply a multidimensional annotation fusion model on annotations at the word level to estimate a parameter which captures relationships between different norms. We then use this parameter at sentence level to estimate the norms. We evaluate our approach by predicting sentence level scores for various normative dimensions and compare with standard word aggregation schemes.

Index Terms: Psycholinguistic normatives, Annotation fusion, Multidimensional annotations.

1. Introduction

Psycholinguistic norms are numeric ratings assigned to linguistic cues such as words or sentences to measure various psycholinguistic constructs. Examples include dimensions such as valence, arousal, and dominance which are used to analyze the affective state of the author of the spoken or written text, along with norms of higher order mental constructs such as concreteness and imageability which have been associated with improvements in learning [1]. The ease of computing the norms has enabled their application in a variety of tasks in natural language processing such as information retrieval [2], sentiment analysis [3], text based personality prediction [4] and opinion mining. The norms are typically annotated at the word level by psychologists who provide numeric scores to a curated list of seed words, which are then extrapolated to a larger vocabulary using either semantic relationships such as synonymy and hyponymy or using word occurrence based contextual similarity [5].

Most applications of psycholinguistic norms in NLP use sentence or document level scores, but manual annotation of the norms at these levels is difficult and not straightforward to generalize. In these cases, estimation of sentence level norms is done by aggregating the word level scores using simple averaging [6, 5], or by using distribution statistics of the word level scores [7]. However, such strategies do not account for the non-trivial dependencies of sentence level semantics on the words, and may not be accurate at estimating the norms at the sentence level. In this work, we propose a new approach to estimate sentence level norms using inferred relationships between different dimensions along with partial annotations of the sentence level norms.

Annotation of the normatives at the sentence level is a challenging task when compared to word level annotations since it involves evaluating the underlying semantics of the sentence in the abstract space of the corresponding dimension, with some dimensions in particular being more difficult than others. For example, imageability, a measure of how easy it is to create a mental image of the input word or sentence, is more difficult to annotate at the sentence level when compared to words. On the other hand, norms such as valence are relatively easier to annotate even at the sentence level in comparison. We use this observation along with the parameters learned from a joint annotation fusion model at word level to predict norms at sentence level.

Annotations are typically performed online using crowdsourcing platforms such as Amazon Mechanical Turk (Mturk), which connect researchers with inexpensive workers from across the globe and provide easy scalability. Annotations are collected from several workers over a large number of instances, often on several related dimensions. These are then combined to obtain estimates for the label of interest, typically using aggregation techniques such as simple averaging or majority voting, or using more nuanced aggregation models which assume a structure for the annotators’ behavior [8]. The annotation dimensions are usually modeled independent of each other, but a few recent publications have explored joint modeling of the dimensions and have highlighted the benefits of this approach [9, 10]. These models assume a joint relationship between the dimensions being annotated, and estimate model parameters that capture this relationship for each annotator, which can be used in estimating the sentence level normatives. Specifically, we can use model parameters learned at the word level to estimate the norms at the sentence level using partial sentence level annotations.

We use the model presented in [10], in which the authors assume a matrix factorization model to capture the annotators’ behavior, in which the annotations are assumed to be based on a linear transformation of the underlying label vector. Parameters of this model include a linear transformation matrix, which captures the individual contributions of each dimension in the annotation output. In our work, we assume that the annotator specific relationships between the dimensions captured by the parameter is comparable at both word and sentence levels. We collect word level annotations on valence, arousal and dominance and train the joint global annotation model from [10] to estimate the annotator parameters including ; we then use the word level estimates for on sentence level ratings from the same set of annotators. To predict sentence level scores of a given normative dimension, we make use of partial annotations on the remaining dimensions along with . Our proposed ap-
proach shows improved performance in predicting the sentence level norms when compared to various word level aggregation strategies.

The rest of the paper is organized as follows. In Section 2, we expand on the joint multidimensional annotation model and detail our data annotation approach in Section 3, followed by experiments in Section 4 and results in Section 5 before concluding in Section 6.

2. Joint multidimensional annotation model

The annotation model is represented in plate notation in Figure 1. In this model, the underlying label vector \( \mathbf{a}_k^m \) for each data instance is defined as a linear regression model as shown in Equation 1. An annotator, indexed by \( k \), is assumed to apply a linear transformation function on vector \( \mathbf{a}_k^m \) to produce the annotation vector \( \mathbf{a}_k^m \) using the matrix \( F_k \) as shown in Equation 2.

\[
\begin{align*}
\mathbf{a}_k^m &= \Theta^T \mathbf{x}_m + \eta_k \\
\mathbf{a}_k^m &= F_k \mathbf{a}_k^m + \eta_k
\end{align*}
\]

where, \( \mathbf{x}_m \in \mathbb{R}^p; \Theta \in \mathbb{R}^{p \times d}; \eta_k \sim N(\mathbf{0}, \sigma^2 I); \sigma^2 \in \mathbb{R}; \eta_k \sim N(\mathbf{0}, \tau_k^2 I); \tau_k^2 \in \mathbb{R} \). \( F_k \in \mathbb{R}^{d \times d} \) is the annotator specific linear transformation matrix. Each annotation dimension value \( a_k^m \) for annotator \( k \) is defined as a weighted average of the vector \( \mathbf{a}_k^m \) with weights given by \( F_k \).

In this model, the feature vector \( \mathbf{x}^m \) corresponding to each instance is assumed to be available, along with the annotations \( \mathbf{a}_k^m \), while the label vectors \( \mathbf{a}^m \) are assumed hidden, as shown in the Figure 1. We use the EM algorithm from [10] to estimate the parameters, listed below for ease of exposition. Detailed derivations for the update equations below can be found in [10].

We use Maximum Likelihood Estimation (MLE) to estimate the model parameters, in which we maximize the model likelihood shown below in Equation 3.

\[
\begin{align*}
\log L &= \sum_{m=1}^{M} \log p(\mathbf{a}_1^m \ldots \mathbf{a}_K^m; F_k, \Theta, \sigma^2, \tau_k^2) \\
&= \sum_{m=1}^{M} \log \int p(\mathbf{a}_1^m \ldots \mathbf{a}_K^m; \mathbf{a}_k^m; F_k, \Theta, \sigma^2) \, d\mathbf{a}_k^m
\end{align*}
\]

Optimizing the above objective is non-trivial due to the presence of the integral within the log function. To address this, we use the well known Expectation Maximization algorithm [11], which uses Jensen’s inequality to derive a lower bound (shown below in Equation 4) on the objective based on current parameter estimates, by computing the expectation with respect to the conditional distribution \( p(\mathbf{a}_k^m|\mathbf{a}_1^m \ldots \mathbf{a}_K^m) \).

\[
\log L = \sum_{m=1}^{M} \mathbb{E}_{\mathbf{a}_k^m|\mathbf{a}_1^m \ldots \mathbf{a}_K^m} \left[ \log \frac{p(\mathbf{a}_1^m \ldots \mathbf{a}_K^m; \mathbf{a}_k^m; F_k, \Theta, \sigma^2)}{q(\mathbf{a}_k^m)} \right] 
\]

This is followed by parameter estimation using maximization. The alternating expectation and maximization steps form the iterations of the EM algorithm.

2.1. EM algorithm

Initialization The model is initialized by assigning the mean of annotations for each data instance as the estimate for \( \mathbf{a}^m \). Given this, the initial parameters are estimated using update equations listed in the maximization step below.

E-step We compute the expected value of \( \mathbf{a}^m \), with respect to the distribution \( p(\mathbf{a}^m|\mathbf{a}_1^m \ldots \mathbf{a}_K^m) \), which is assumed to be it’s soft estimate for each data instance.

\[
\begin{align*}
\mathbb{E}_{\mathbf{a}_k^m|\mathbf{a}_1^m \ldots \mathbf{a}_K^m} \left[ \mathbf{a}_k^m \right] &= \Theta^T \mathbf{x}_m + \sigma^2 \mathbf{a}_k^m \sum_{i=1}^{K} \mathbf{a}_k^m \mathbf{a}_k^m^T \sum_{i=1}^{K} \mathbf{a}_k^m \mathbf{a}_k^m^T \mathbf{a}_k^m - \mu_k^m \sum_{i=1}^{K} \mathbf{a}_k^m \mathbf{a}_k^m^T \mathbf{a}_k^m - \sigma^2 \sum_{i=1}^{K} \mathbf{a}_k^m \mathbf{a}_k^m^T \mathbf{a}_k^m - \sigma^2 \sum_{i=1}^{K} \mathbf{a}_k^m \mathbf{a}_k^m^T \mathbf{a}_k^m ^T \mathbf{a}_k^m - \sigma^2 \sum_{i=1}^{K} \mathbf{a}_k^m \mathbf{a}_k^m^T \mathbf{a}_k^m
\end{align*}
\]

M-step Given the soft estimate for \( \mathbf{a}^m \), parameter estimates are computed by maximizing Equation 4. The update equations for this step are listed below.

\[
\begin{align*}
\Theta &= (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbb{E}[\mathbf{a}^m]) \\
F_k &= \left( \sum_{m=1}^{M} \mathbf{a}_k^m \mathbf{a}_k^m^T \right) \left( \sum_{m=1}^{M} \mathbb{E}[\mathbf{a}_k^m (\mathbf{a}_k^m)^T] \right)^{-1} \\
\sigma^2 &= \frac{1}{md} \sum_{m=1}^{M} \left[ \mathbb{E}[(\mathbf{a}_k^m)^T \mathbf{a}_k^m] - 2 \mathbb{E}[(\Theta^T \mathbf{x}_m) \mathbb{E}[\mathbf{a}_k^m (\mathbf{a}_k^m)^T]] + \mathbb{E}[(\mathbf{x}_m^T \Theta \Theta^T \mathbf{x}_m)] \right] \\
\tau_k^2 &= \frac{1}{md} \sum_{m=1}^{M} \left[ \mathbb{E}[(\mathbf{a}_k^m)^T \mathbf{a}_k^m] - 2 \mathbb{E}[(F_k^T \mathbf{a}_k^m \mathbf{a}_k^m^T (\mathbf{a}_k^m)^T)] + \mathbb{E}[(F_k^T F_k^T \mathbf{a}_k^m (\mathbf{a}_k^m)^T)] \right]
\end{align*}
\]

Termination We terminate the algorithm when the change in model log-likelihood reduces to less than 0.001% from the previous iteration.

3. Data annotation

We performed two sets of experiments, collecting word and sentence level annotations on specific dimensions in each. In the first experiment (which we refer to as VAD from now), we collected annotations on the affective norms of Valence, Arousal and Dominance using Murk for words sampled from [12]. This corpus was chosen because it provides expert ratings on Valence, Arousal and Dominance for nearly 14,000 English words. Annotators were asked to provide numeric ratings between 1 to
Sentence level annotations were collected on sentences from the Emobank corpus [13], which includes expert ratings on valence, arousal and dominance. For 10000 English sentences, we collected word and sentence level annotations on arousal and dominance along with the remaining norms. For example, in the VAD experiment, while predicting sentence level scores of valence, we use the sentence level annotator ratings on arousal and dominance along with the word level parameter matrix \( F_{\text{word}}^k \). The use of partial annotations enables us to predict sentence level norms on challenging psycholinguistic dimensions using ratings on dimensions which maybe easier to annotate.

Given annotator parameters \( F_k \) estimated at the word level, we use partial annotator ratings at the sentence level to predict the remaining norms. For example, in the VAD experiment, while predicting sentence level scores of valence, we use the sentence level annotator ratings on arousal and dominance along with the word level parameter matrix \( F_{\text{word}}^k \). The use of partial annotations enables us to predict sentence level norms on challenging psycholinguistic dimensions using ratings on dimensions which maybe easier to annotate.

\[
\begin{bmatrix}
    a_1^{m,d'} \\
    \vdots \\
    a_K^{m,d'}
\end{bmatrix} =
\begin{bmatrix}
    F_1^\text{word}[d',:] \\
    \vdots \\
    F_K^\text{word}[d',:]
\end{bmatrix}
\begin{bmatrix}
    a_k^m
\end{bmatrix}
\]

(5)

In our second experiment (which we refer to as IGP from now), we collected word and sentence level annotations on three new psycholinguistic normative dimensions: imageability, which measures the degree of the stimulus’ proclivity to create a mental picture; genderladenness, which measures the degree of masculine or feminine association evoked by the stimulus; and pleasantness, which measures the degree of pleasant feelings associated with the stimulus. We used the same words and sentences used in our previous experiment for annotations on valence, arousal and dominance. Since we do not have expert ratings for pleasantness, imageability and genderladenness, we use training error as a proxy, where we train linear regression models on labels predicted by the different schemes and compare their MSE to evaluate label performance, since low MSE can be indicative of higher correlations between the labels and features.

4. Experiments

...
where, \( d' \in \{1 \ldots D \} ; d' \neq d \) is the dimension to predict. \( \mathbf{a}^m_\mathbf{w} \) is estimated using linear regression.

In both our experiments, we make use of the IID Gaussian noise assumption in Equation 2, which reduces the task of predicting the sentence level norm to a linear regression problem shown in Equation 5. Rows of the matrix \( F_{\mathbf{w} \mathbf{d}} \) are treated as features of the regression model with vector \( \mathbf{a}^m_\mathbf{w} \) as the regression parameter. Given sentence level partial annotations \( \mathbf{a}^m_{\mathbf{w,d}} \) (vector \( \mathbf{a}^m_\mathbf{w} \) with \( \mathbf{a}^m_{\mathbf{d}} \) removed), and matrix \( F_{\mathbf{w} \mathbf{d}} \), the regression parameter vector \( \mathbf{a}^m_\mathbf{w} \) can be estimated using normal equations or gradient descent. For each dimension within a given experiment, we use Equation 5 to estimate the sentence level norms. The features \( \mathbf{x}_m \) used in both our experiments were 300 dimensional GloVe embeddings [14] at word level, which were aggregated using simple averaging at sentence level.

In the VAD experiment, we compare the predicted dimensions with expert ratings from the Emobank corpus, which acts as our reference to evaluate model performance. For baselines, we compute different aggregations of word level normative scores after filtering out non-content words as is common in literature [6]. Word level scores for the norms were computed using the approach described in [5]. We used unweighted average, maximum, minimum and sum of the word level norms as the baseline aggregation functions.

In the IGP experiment, we train linear regression models using predictions from the proposed model and directly compare the training set error with baselines. Low training error implies higher learnability (due to better correlations with the features) of the predicted signal and serves as a crude proxy for quality. For baselines, we use training error from labels obtained by simple averaging of word level normative scores, and sentence level average of annotations.

We use Concordance Correlation Coefficient (\( \rho_c \)) [15] and the Pearson’s correlation coefficient (\( \rho \)) as evaluation metrics. \( \rho_u \) measures any departures from the concordance line (line passing through the origin at 45° angle). Hence it is sensitive to rotations or rescaling in the predicted values of \( \mathbf{a}^m_\mathbf{w} \). Given two samples \( x \) and \( y \), the sample concordance coefficient \( \hat{\rho}_c \) is defined as shown below.

\[
\hat{\rho}_c = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2}
\]  

where \( \bar{x} \) and \( \bar{y} \) are sample means, \( s_x \) and \( s_y \) are sample standard deviations, and \( s_{xy} \) is the sample covariance.

5. Results

5.1. VAD

Figure 2 shows the performance of the proposed model along with the different baselines. As seen from the figure, the proposed model outperforms the baselines in predicting valence and arousal in both evaluation metrics, suggesting the efficacy of the approach. Using partial ratings at sentence level along with matrix \( F_k \) which captures relationships between the dimensions, the proposed approach seems to outperform the baseline word aggregation schemes in these two dimensions. Performance in \( \rho_c \) appears to be lower than \( \rho_u \) suggesting the presence of a rotation in the predicted values. This can be attributed to the unidentifiability commonly observed in matrix factorization models such as the annotation fusion model of [10]. Common solutions to address this involve assuming a suitable prior on the parameter \( F_k \), which may lead to better estimates of \( \rho_c \).

Model performance on dominance, on the other hand, is considerably low in both metrics. To further investigate the reason for this, we examined the performance of the best possible annotator for each dimension in this experiment and compare their predictions with the expert ratings from the Emobank corpus in Figure 3. Evidently, for dominance, we notice very low correlation and high MSE between our best annotators and the experts, suggesting a high disagreement for this dimension. This may have been due to a possibly differing definition and/or interpretation of dominance between the two sets of annotators.

5.2. IGP

In our second experiment, we use model training error as a proxy for evaluating prediction quality since we do not have expert ratings. Figure 4 shows the training error for the proposed model when compared with two baselines. The proposed model shows lowest training error in predicting imageability while the performance is relatively worse in gender ladenness and pleasantness, suggesting relatively stronger dependency of imageability on the other dimensions.

6. Conclusion

We presented a novel computational approach to estimate sentence level psycholinguistic norms using joint multidimensional annotation fusion. We evaluate our approach by predicting sentence level normatives on various dimensions in two different experiments, and showed improvements in specific cases. Future work includes evaluating the model on more abstract dimensions such as concreteness.
7. References


