Recursive Neural Networks for Coding Therapist and Patient Behavior in Motivational Interviewing

Michael Tanana[1], **Kevin Hallgren**[2], **Zac Imel**[1], **David Atkins**[2], **Padhraic Smyth**[3], **and Vivek Srikumar**[4]
[1] Department of Educational Psychology, University of Utah, Salt Lake City, UT
[2] Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington
[3] Department of Computer Science, University of California, Irvine, CA
[4] School of Computing, University of Utah, Salt Lake City, UT
michael.tanana@utah.edu khallgre@uw.edu zac.imel@utah.edu
datkins@u.washington.edu smyth@ics.uci.edu svivek@cs.utah.edu

Abstract

Motivational Interviewing (MI) is an efficacious treatment for substance use disorders and other problem behaviors (Lundahl and Burke, 2009). However, little is known about the specific mechanisms that drive therapeutic change. A growing body of research has focused on coding within-session language to better understand how therapist and patient language mutually influence each other and predict successful (or unsuccessful) treatment outcomes. These studies typically use human raters, requiring considerable financial, time, and training costs for conducting such research. This paper describes the development and testing of a recursive neural network (RNN) model for rating 78,977 therapist and patient talk turns across 356 MI sessions. We assessed the accuracy of RNNs in predicting human ratings for client speech and compared them to standard n-gram models. The RNN model showed improvement over ngram models for some codes, but overall, all of the models performed well below human reliability, demonstrating the difficulty of the task.

1 Introduction

1.1 Motivational Interviewing

Motivational Interviewing (MI) (Miller and Rollnick, 2012) is a counseling style that attempts to highlight and resolve patient ambivalence about behavioral change. To achieve these aims, MI theory emphasizes that therapists should use specific MI-consistent strategies, such as fostering collaboration rather than confrontation, emphasizing patient autonomy rather than therapist authority, and eliciting discussion of the factors that motivate patients to change or not change their behavior. During MI sessions, therapists are instructed to attend to patient *change talk* (i.e., language that indicates a desire, reason, or commitment to make a behavioral change), and *sustain talk* (i.e., language that indicates a desire, reason, or commitment against making a behavioral change). Therapists are further instructed to respond to such change and sustain talk in specific, MI-consistent manners. For example, therapists are instructed to frequently use open questions and to reflect patient language with the goal of eliciting change talk from patients. Likewise, therapists are instructed to minimize their use of behaviors such as confrontation, warning, and giving advice without permission.

MI researchers have developed several coding systems for identifying these types of patient and therapist language. The information provided by these MISC ratings often provides critical data for a variety of research and training purposes. For example, such coding data can be used to assess therapists' fidelity to using MI (e.g., based on the amount of MI-consistent and MI-inconsistent therapist behaviors), to understand the temporal relationships between therapist and patient behaviors (e.g., through sequential analysis of therapist and patient codes), or to understand how in-session behaviors predict out-of-session behavioral change (e.g., therapist and patient language predicting reductions in substance use). These coding systems typically require human coders to listen to psychotherapy sessions and manually label each therapist and patient utterance using codes derived from MI theory. For example, one of the most versatile but time consuming coding systems, the Motivation Interview-

Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 71–79,
Denver, Colorado, June 5, 2015. ©2015 Association for Computational Linguistics

ing Skill Code (Houck et al., 2012) assigns codes to every therapist and patient utterance (defined as a single idea within a section of speech) using over 30 different predefined codes (See examples below in Figure 1).

> **Counselor**: “How do you feel about your progress so far?” **(Open Question)**
> **Patient**: “Everyone’s getting on me about my drinking.” **(Follow-Neutral)**
> **Counselor**: ”Kind of like a bunch of crows pecking at you.” **(Complex Reflection)**
> **Patient**: “I’m not sure I can finish treatment.” **(Sustain Talk)**
> **Counselor**: “You’re not sure if you can finish treatment.” **(Simple Reflection)**
> **Patient**: “I drank a couple of times this week when I was with my brother **(Sustain Talk)**. I want to quit so badly **(Change Talk)**, but I don’t think I can do it.” **(Sustain Talk)**

Figure 1: Example of MISC codes from (Houck et al., 2012)

1.2 Machine Learning and Psychotherapy Coding

There are few studies that have used machine learning to assess therapist and patient behavior in psychotherapy sessions. Most of these methods have relied heavily on n-grams (i.e., specific words or phrases) and have used a bag of words approach where the temporal ordering of n-grams within an utterance is mostly ignored, thereby losing information about the functional relationships between words.

For example (Atkins et al., 2014) used topic modeling to predict utterance-level MISC codes in 148 MI sessions obtained from studies of primary care providers in public safety net hospitals and brief interventions for college student drinking. The topic models were able to predict human ratings of utterances with high accuracy for many codes, such as open and closed questions or simple and complex reflections (Cohen’s kappa all >0.50). However, the topic models struggled to accurately predict other codes, such as patient *change talk* and *sustain talk* (Cohen’s kappa all <0.25). The limitations in the prediction model were attributed to multiple sources, including low inter-rater agreement among the human raters, the limited information provided within the relatively small number of n-grams contained in single utterances, the inability to incorporate the local context of the conversation in the predictive model, and the lack of a uniform linguistic style associated with some codes (e.g., questions typically contain keywords such as “what” or “how”, but *change talk* does not).

Using a subset of the same data, (Can et al., 2012) used multiple linguistic features to predict utterance-level therapist reflections with reasonably high accuracy, F1 = 0.80. Specifically, Can et al. used N-grams (i.e., specific words and phrases), similarity features (i.e., overlapping N-grams between therapist utterances and patient utterances that preceded), and contextual meta-features (i.e., words in the surrounding text) with a maximum-entropy Markov model and found improved performance relative to models that did not include similarity or meta-features. However, this study did not test the prediction of language categories that were difficult to predict in Atkins et al., such as *change talk* and *sustain talk*.

1.3 Aims

An important problem with the word and n-gram based models is that they do not account for syntactic and semantic properties of the text. In this work, we study the question of using dense vector features and their compositions to address this issue

To our awareness, no research to date has tested the use of recursive neural networks (RNNs) for predicting MISC codes. It is possible that a model capturing semantic and syntactic similarity in text can perform better than n-gram models in identifying reflections in MI sessions. The present study aimed to test (1) whether recursive neural networks (RNNs) (Socher, 2014) can be used to predict utterance-level patient MISC codes and (2) whether RNNs can improve the prediction accuracy of these codes over n-gram models.

Following the basic procedure described in (Socher, 2014), we developed a Recursive Neural

Network model to achieve these aims. We used the Stanford parser (Klein and Manning, 2003) to create parse trees that modeled the language structure of patient and therapist utterances. These sentence-level models were then used as input into a Maximum Entropy Markov Model (MEMM), a type of sequence model that uses the sentence and surrounding context to predict MISC codes. The recursive neural networks were designed using the 'standard' model (Socher et al., 2011) with a single weight matrix to combine each node in the tree.

We tested both a standard RNN model and an RNN that utilized a dependency parsing of the sentence. Once a final model was tuned, the performance of each model predicting *change talk* and *sustain talk* codes was examined by comparing RNNs with an n-gram based model using cross-validation.

The main goals of this paper are to

1. Define the challenging and interesting problem of identifying client *change* and *sustain talk* in psychotherapy transcripts.

2. Explore and evaluate methods of using continuous word representations to identify these types of utterances and

3. Propose future directions for improving the performance of these models

2 Data

We used the dataset constructed as part of a collaborative project between psychologists at the University of Utah and the University of Washington and computer scientists and engineers at the University of California, Irvine and University of Southern California. The dataset consists of 356 psychotherapy sessions from 6 different studies of MI, including the 5 studies (148 sessions) reported in (Atkins et al., 2014). The original studies were designed to assess the effectiveness of MI at a public safetynet hospital (Roy-Byrne et al., 2014), the efficacy of training clinicians in using MI (Baer et al., 2009), and the efficacy of using MI to reduce college student drinking (Tollison et al., 2008; Neighbors et al., 2012; Lee et al., 2013; Lee et al., 2014). All sessions have utterance level MISC ratings totaling near 268,000 utterances in 78,977 talk turns. A subset of sessions was coded by multiple raters to estimate inter-rater reliability, which serves as a theoretical lower-bound for the predictive performance.

3 Modeling MISC Codes

3.1 Sequence Labeling

All of the models attempted to correctly label utterances as a single sequence. For example, a patient may speak two or three times in a row, then the therapist may speak once. Each utterance code is predicted by the preceding utterance label, regardless of the speaker. Both patient and therapist utterances were combined into this sequence model.

All sequence models were Maximum Entropy Markov Models (MEMM)(McCallum and Freitag, 2000). At test time, the sequences for the codes were inferred using the Viterbi algorithm. The models all differed in their feature inputs into the MEMM. The N-gram model used sparse input vectors representing the presence of the various unigrams, bigrams and trigrams in each utterance. The RNN models used the final sentence vector as the input into the MEMM model. The RNN models were allowed to learn from the mistakes in the MEMM models through backpropogation.

Even though the purpose of this model was to predict patient change and sustain talk, we attempted to predict all codes in the sequence to assist in the task due to the relationship between change talk, sustain talk, and other MISC codes. Other codes identified by the models included reflect, affirm, giving information, facilitate, open questions, closed questions, advise, confront, and follow-neutral (See (Houck et al., 2012)).

It should be noted that the MEMM models only used the previous utterance codes (or predicted codes) and the current utterance for feature inputs. We were attempting in this study to identify the best sentence model. At a later point in time, similar work will be done testing various iterations of sequence models to find the optimal version, after the best sentence level model has been chosen. One of the reasons for choosing a MEMM over a conditional random field was to allow for joint training of the RNN models and the sequence model (with a MEMM, it is easy to backpropogate errors from the sequence model to the sentence model).

3.2 Word Based Features

Our first feature set for utterances is defined using indicators for n-grams. In all cases, the speaker of the utterance (patient or therapist) was considered to be known. That is, the models only had to distinguish between codes applicable for each speaker role and did not have to distinguish the roles of the speakers as patient or therapist. We trained two different models – one that uses indicators for only unigrams in the utterance and the second that uses indicators for unigrams, bigrams and trigrams in the utterance.

3.3 Recursive Neural Network

Our second feature set uses recursive neural network (RNN) models, which are variants of the ideas presented in (Socher, 2014). The models were initialized with word vectors (i.e., numeric representations of word tokens) that were pre-trained using word vectors generated by the Glove model (Pennington et al., 2014). The RNNs in this paper relied mostly on the standard model for combining nodes of a recursive tree. For example, for combining word vector 1 a_1 (e.g., numeric representation of ”hate”) and word vector 2 a_2 (e.g., numeric representation of “hangovers”), the two vectors are multiplied through a weight matrix W_m that is shared across the tree in order to combine the individual words (e.g., “hate” and “hangovers”) into a new vector that combines the meaning of both inputs, $a_{1,2}$ (e.g., “hate hangovers”). This is performed through the function:

$$p_{1,2} = tanh\left(W_m \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + b\right)$$

where a_1,a_2 and $p_{1,2}$ are all $\mathbb{R}^{dx1}$ where d is dimensionality value for the word vectors that is chosen by the researcher. Typically, several sizes of word vectors are tried to discover the optimal length for different types of problems. Based on cross-validated comparisons of different vector lengths, 50 dimensional word vectors were found to have the best overall performance and were used in the present study. Importantly, the non-linearity of hypertangent is used, which constrains the outputs to be between -1 and +1.

The top level vector of the RNN, which represents the whole linguistic utterance, was used as input into a MEMM to combine individual utterances with information from the surrounding linguistic context.

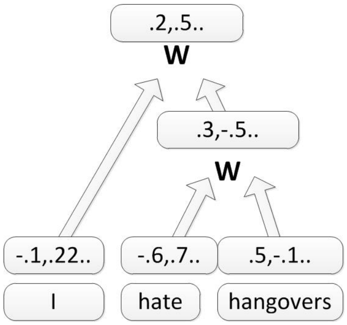


Figure 2: RNN Model. Each level of the parse tree is represented by a vector of 50 numeric values. Higher-level phrases and subphrases are modeled by multiplying the child node vectors through a weight matrix W_m.

The learning utilized backpropagation through structure (Goller and Kuchler, 1996). In other words, errors made at the top of the tree structure gave information that allowed the parameters lower in the model to learn, improving prediction accuracy. Weight updates were performed using adagrad with the diagonal variant (see Technical Appendix)(Duchi et al., 2011). The advantage of this weight update method is that it allows the model to learn faster for more rare words and to learn more slowly for frequently seen words.

3.4 Dependency RNN

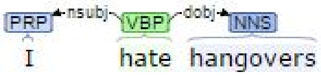


Figure 3: Example Dependency Parse

The final feature vector we tested was based on (Socher et al., 2014), with some important differences. In our model, we used the dependency tree from the Stanford parser to create a collection of edges, each with its label. For example, in figure 3 the dependency parse can be thought of as having three node with two labeled edges. The edge between “I” and “hate” has the label *nsubj* for nominal subject. In our dependency RNN we multiply the word vectors for “I” (the child node) and “Hate” (the parent node) through a weight matrix that is specific to the label *nominal subject*.

The model cycles through all of the edges in the dependency tree, then sums the output vectors. After summing, a hypertangent nonlinearity is applied to get the final feature vector for the sentence. Formally, this can be written as follows:

$$p_s = tanh\left(\sum\nolimits_{(p,c,\ell)\in D(s)} W_\ell \begin{bmatrix} a_p \\ a_c \end{bmatrix} + b\right)$$

a_p is the parent word vector and a_c is the child word vector. In this case, W_ℓ is the weight matrix specific to the dependency relationship for that specific label. The model sums over all parent p, child c and label ℓ triads in the dependency parse of the sentence ($D(s)$) and then adds an intercept vector b. The weight matrix is initialized to the shared weight matrix from the pre-trained standard RNN, but then is allowed to learn through backpropagation. The final model combines the output of the standard RNN and the dependency RNN as adjacent vectors. Both models share their word vectors and learn these vectors jointly.

4 Evaluation

To evaluate the performance of the RNN and n-gram models we compared precision (i.e., proportion of model-derived codes that matched human raters), recall (i.e., proportion of human-rated codes that were correctly identified by the model), and F1 scores (i.e., the harmonic mean of precision and recall) for each model. The current results are an early stage in the process toward developing a final model. As such, all models were evaluated using 5 fold cross validation on the section of the dataset that is designated as training data (which is two thirds of the total dataset). The cross validation subsets were divided by session (so each session could only occur in one or the other subsets). The testing section of the data will be used at a later date when the modeling process is complete.

5 Results

5.1 Prediction Accuracy

When predicting change talk (see table 1), the models varied in their performance. Unigram-only and unigram, bigram, and trigram models had F1 scores of 0.17 and 0.21, respectively. The standard RNN had a much lower F1 score of 0.06. The dependency RNN outperformed both the n-gram models and the standard RNN on F1 score (0.22). While the dependency RNN performed best on F1 score and precision, the Uni,Bi and tri-gram model tied for recall of change talk. These values were all relatively low compared to the theoretical upper bound of predictive performance based on the estimated human agreement, F1 = 0.61.

Table 1: Cross Validation Results: Change Talk

Model	Precision	Recall	F1
Uni-gram	.24	.13	.17
Uni,Bi,Tri-Gram	.28	**.18**	.21
Standard RNN	.15	.03	.06
Dependency RNN	**.29**	**.18**	**.22**
Human Agreement	.73	.42	.61

Table 2: Cross Validation Results: Sustain Talk

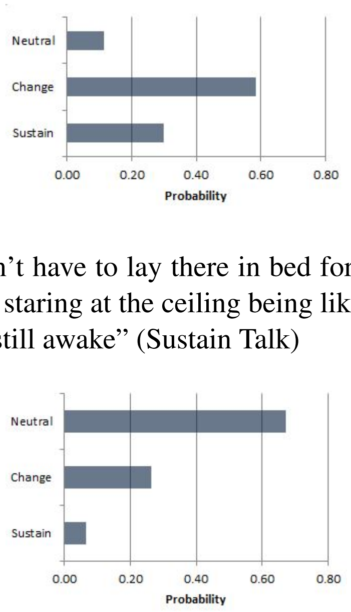
Model	Precision	Recall	F1
Uni-gram	.26	.20	.22
Uni,Bi,Tri-Gram	**.33**	.20	**.24**
Standard RNN	.19	**.23**	.21
Dependency RNN	.26	.19	.22
Human Agreement	.66	.53	.59

When predicting sustain talk (table 2), the unigram model, unigram, bigram, and trigram model, and the standard RNN all performed similarly in terms of F1 scores (F1 = 0.21 to 0.24), with the Uni, Bi and trigram model performing the best (.24). The Standard RNN had the highest recall (.23), but had the lowest precision (.19) As with change talk, all models had relatively low F1 scores compared to the F1 scores between human raters, F1 = 0.59.

5.2 Examples

Figure 4 shows two example sentences from the test sample of the dataset, one which was predicted correctly from the dependency RNN and one that was predicted incorrectly. Below each sentence is a chart with the predicted probability that it was change talk, sustain talk or follow-neutral (i.e., neither change talk or sustain talk). In the first example, the dependency RNN did well at identifying a simple change statement. Similarly simple utterances, such as “I don’t want to drink anymore” or “I enjoy drinking alcohol” were typically coded correctly as change talk or sustain talk. But more complicated utterances, like the second example in figure 4 were less likely to be coded correctly. (Note that the second utterance depends more than the context of previous statements in the conversation, which involved the patient discussing reasons for smoking marijuana.)

”Because I don’t really want to have to smoke more” (Change Talk)



“I don’t have to lay there in bed for three hours staring at the ceiling being like why am I still awake” (Sustain Talk)

Figure 4: Example Codings

6 Conclusions

In general, predicting change and sustain talk is a non-trivial task for machine learning. It involves a subtle understanding of the context of a phrase and involves more than just the words in a single sentence. These early models are able to correctly identifying many statements as change talk or sustain talk, particularly for sentences with simple structures such as “I want to stop drinking”. However, these models appear have a harder time with sentences that are longer and have greater complexity and sentences that require more contextual information based on previous statements. These initial results show that our dependency RNN has the ability to outperform n-gram models on identifying client change talk, but this performance gain did not apply to sustain talk.

As shown in (Can et al., 2012) and (Atkins et al., 2014), machine learning techniques are able to reliably identify important linguistic features in MI. This study represents an initial attempt at predicting the more difficult-to-identify patient behaviors, which are central to much of the research on MI. More work is needed to improve these models, and it is likely that performance could be improved by going beyond word counting models, for example, by using the syntactic structure of sentences as well as the context of surrounding utterances.

NLP applications have been successful in areas in which human annotators can clearly label the construct of interest (e.g., sentiment in movie reviews(Socher et al., 2013b), classifying news articles(Rubin et al., 2012)). Psychotherapy generally and ‘change talk’ within MI specifically are often concerned with latent psychological states of human experience. Verbalizations of reducing drug use are hypothesized to be observed indicators of a patient’s inclination to change their behavior and is mutually dependent on both their own previous linguistic behavior as well as the therapist’s. This is a challenging, new arena for NLP application and development, and one that will only be successful through the tight collaboration of NLP researchers and domain experts.

6.1 Limitations

There were some important limitations to this initial study. First, we have not yet systematically explored all of the possible options for discrete word models. For example, one could use the dependency tree to create non-sequential n-grams that capture longer range dependencies than traditional n-grams. We acknowledge that part of the advantage given to the RNN is the information the dependency tree provides and that it is possible for discrete word models to use this type of information as well. Second, not all of the possible combinations of word dimensions and word models were tried. Because of limitations in available compute capacity, only promising combinations were tested. Third, there was a moderate degree of disagreement between human raters. These human ratings were required for training each method and were used as the criterion for classifying correct or incorrect ratings, and error in these ratings limits the performance of the models.

References

David C Atkins, Mark Steyvers, Zac E Imel, and Padhraic Smyth. 2014. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation science : IS*, 9(1):49, January.

John S. Baer, Elizabeth a. Wells, David B. Rosengren, Bryan Hartzler, Blair Beadnell, and Chris Dunn. 2009. Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors. *Journal of Substance Abuse Treatment*, 37(2):191–202.

Leon Bottou. 2014. From Machine Learning to Machine Reasoning. *Machine Learning*, 94(2):133–149.

Dogan Can, Panayiotis G. Georgiou, David C Atkins, and Shrikanth Narayanan. 2012. A Case Study: Detecting Counselor Reflections in Psychotherapy for Addictions using Linguistic Features. In *Proceedings of InterSpeech*.

John Duchi, E Hazan, and Y Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning . . .*, pages 1–40.

Christoph Goller and Andreas Kuchler. 1996. Learning task-dependent distributed representations by backpropagation through structure. *Proceedings of International Conference on Neural Networks (ICNN'96)*, 1.

Jon. M. Houck, TheresaB. Moyers, William R. Miller, Lisa. H. Glynn, and Kevin. A. Hallgren. 2012. ELICIT Motivational Interviewing Skill Code (MISC) 2.5 coding manual. Technical report, Unpublished coding manual, University of New Mexico.

Dan Klein and CD Manning. 2003. Accurate unlexicalized parsing. In *ACL '03 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 423–430.

Christine M Lee, Jason R Kilmer, Clayton Neighbors, David C Atkins, Cheng Zheng, Denise D Walker, and Mary E Larimer. 2013. Indicated Prevention for College Student Marijuana Use: A Randomized Controlled Trial. *Journal of consulting and clinical psychology*, 81(4):702–709.

Christine M Lee, Clayton Neighbors, Melissa a Lewis, Debra Kaysen, Angela Mittmann, Irene M Geisner, David C Atkins, Cheng Zheng, Lisa a Garberson, Jason R Kilmer, and Mary E Larimer. 2014. Randomized controlled trial of a Spring Break intervention to reduce high-risk drinking. *Journal of consulting and clinical psychology*, 82(2):189–201.

Brad W. Lundahl and Brian L. Burke. 2009. The effectiveness and applicability of motivational interviewing: A practice-friendly review of four meta-analyses.

Andrew McCallum and Dayne Freitag. 2000. Maximum entropy markov models for information extraction and segmentation. *Proceedings of the 17th International Conference on Machine Learning*, pages 591–598.

William R. Miller and Stephen Rollnick. 2012. *Motivational Interviewing, Third Edition: Helping People Change*. The Guilford Press, New York, NY, third edit edition.

Clayton Neighbors, Christine M. Lee, David C. Atkins, Melissa a. Lewis, Debra Kaysen, Angela Mittmann, Nicole Fossos, Irene M. Geisner, Cheng Zheng, and Mary E. Larimer. 2012. A randomized controlled trial of event-specific prevention strategies for reducing problematic drinking associated with 21st birthday celebrations. *Journal of Consulting and Clinical Psychology*, 80(5):850–862.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Peter Roy-Byrne, Kristin Bumgardner, Antoinette Krupski, Chris Dunn, Richard Ries, Dennis Donovan, Imara I. West, Charles Maynard, David C. Atkins, Meredith C. Graves, Jutta M. Joesch, and Gary a. Zarkin. 2014. Brief Intervention for Problem Drug Use in Safety-Net Primary Care Settings. *Jama*, 312(5):492.

Timothy N. Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. 2012. Statistical topic models for multi-label document classification. *Machine Learning*, 88:157–208.

Richard Socher, Jeffrey Pennington, and EH Huang. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. *Proceedings of the EMNLP*, (ii):151–161.

Richard Socher, John Bauer, CD Manning, and AY Ng. 2013a. Parsing with compositional vector grammars. In *Proceedings of the ACL conference*.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and C. Potts. 2013b. Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the Conference on Empirical Methods*.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded

Compositional Semantics for Finding and Describing Images with Sentences. *Transactions of the Association for Computational Linguistics (TACL)*, 2:207–218.

Richard Socher. 2014. *Richard Socher August 2014*. Dissertation, Stanford.

Sean J. Tollison, Christine M. Lee, Clayton Neighbors, Teryl a. Neil, Nichole D. Olson, and Mary E. Larimer. 2008. Questions and Reflections: The Use of Motivational Interviewing Microskills in a Peer-Led Brief Alcohol Intervention for College Students. *Behavior Therapy*, 39:183–194.

7 Technical Appendix

7.1 General Details on Learning

The loss function for all outputs o was: $E = \frac{1}{2}\sum_{o}(t_o - y_o)^2$. All of the weights were given random initialization between -.001 and .001. The selection of all hyper-parameters including ada-grad learning rate, weight decay and word vector size were chosen using 5 fold cross validation in the training set of data. As mentioned in the paper, these results will be tested on a third of the data reserved for testing at a later stage in this process when we have selected a final set of models. The optimal learning rate for the RNN models was .1, and the optimal weight decay was 1×10^{-7}. It should be noted that selection of hyperparameters had a major impact on the success of the RNN models. Different learning rates would often result in RNN models that performed half as well as the optimal models.

All models were pre-trained on the general psych corpus (The corpus is maintained and updated by the Alexander Street Press (http://alexanderstreet.com/) and made available via library subscription) using an idea from (Bottou, 2014) called a corrupted frame classifier. The idea is to try to get the model to predict which parts of its parse tree are 'real' sentences and which ones are 'corrupted', that is, one word has been replaced with a random word. Early testing found this unsupervised pre-training significantly improves the performance of the final models.

7.2 Ada-Grad

The training for both the Recursive Neural Nets and the Maximum Entropy Markov Models in this paper utilize stochastic gradient descent, based on common conventions in the machine learning literature, with one important exception. We used the adaptive gradient descent algorithm to adjust our learning rate (Duchi et al., 2011). We opted to use the diagonal variant for simplicity and conservation of memory. The Ada-grad variant to stochastic gradient descent, basically adapts the change in the gradient so that parameters that have many updates will update more slowly over time. Whereas, the parameters that have very few updates will make larger changes. It is obvious that this is advantageous given the fact that in RNN's, the main weight parameters might update on every case, whereas certain word vectors may only have a couple of presentation of an entire corpus. The classic weight update for stochastic gradient descent is $\theta_{t+1} = \theta_t - \alpha G_t$ Where θ are the weights that are being estimated and α is the learning rate. G_t is the gradient at time t. For Ada-grad, we just need to save a running total of the squared gradient, elementwise (we call it γ here):

$$\gamma_t = \gamma_{t-1} + G_t^2$$

And then we add an adjustment to the update step (again, elementwise). Divide the gradient by the square root of the running total sum of squared gradients:

$$\theta_{t+1} = \theta_t - \alpha_t \frac{G_t}{\sqrt{\gamma_t} + \beta}$$

Where β is a constant.

7.3 Notes on Code

Most of the code for this project was written specifically for this problem, but some additional libraries were used. All matrix operations used the Universal Java Matrix Package: http://sourceforge.net/projects/ujmp/files/. Some spell checking was required of some of the data and the open source Suggester was used: http://www.softcorporation.com/products/suggester/ . Version 3.2 of the Stanford parser was used to create the parse trees for the RNN's. (Klein and Manning, 2003; Socher et al., 2013a)