Couples Behavior Modeling and Annotation Using Low-Resource LSTM Language Models

Shao-Yen Tseng[1], *Sandeep Nallan Chakravarthula*[1], *Brian Baucom*[2], *Panayiotis Georgiou*[1]

[1]University of Southern California, Department of Electrical Engineering, USA
[2]The University of Utah, Department of Psychology, USA

{shaoyent,nallanch}@usc.edu, brian.baucom@utah.edu, georgiou@sipi.usc.edu

Abstract

Observational studies on couple interactions are often based on manual annotations of a set of behavior codes. Such annotations are expensive, time-consuming, and often suffer from low inter-annotator agreement. In previous studies it has been shown that the lexical channels contain sufficient information for capturing behavior and predicting the interaction labels, and various automated processes using language models have been proposed. However, current methods are restricted to a small context window due to the difficulty of training language models with limited data as well as the lack of frame-level labels. In this paper we investigate the application of recurrent neural networks for capturing behavior trajectories through larger context windows. We solve the issue of data sparsity and improve robustness by introducing out-of-domain knowledge through pretrained word representations. Finally, we show that our system can accurately estimate true rating values of couples interactions using a fusion of the frame-level behavior trajectories. The ratings predicted by our proposed system achieve inter-annotator agreements comparable to those of trained human annotators.

Importantly, our system promises robust handling of out of domain data, exploitation of longer context, on-line feedback with continuous labels and easy fusion with other modalities.

Index Terms: Behavioral Signal Processing, Dyadic Conversational Language, Recurrent Neural Network, LSTM

1. Introduction

Observational studies in psychotherapy are based on the evaluation of human behaviors that are exhibited in interactions between patients. In couples therapy, for example, behaviors in couple interactions are identified and annotated across numerous dimensions, such as negativity, blame, or humor. Annotation by humans of these dimensions is a time consuming and expensive task. Manual annotation of interaction sessions first requires the training of human annotators according to a detailed coding manual [1, 2]. Annotators have to be trained to give ratings in a consistent manner across all sessions, unbiased by personal influences. Evaluations are then performed on the annotators to select those with the highest agreements for actual annotation tasks. Even then, disagreement in human annotations is inevitable [2].

Estimating behavior is a complex task. Human behavior manifests over longer time frames than emotions and requires a larger context window to be identified correctly. In addition, many different dimensions of behavior have only subtle differences between them which makes tracking the presence of a specific behavior over a long time range more difficult.

This work is supported by NSF and DoD.

Recently there have been great advancements in integrating Signal Processing and Natural Language Processing methods for modeling behavior states and identifying behavioral cues in human interactions [3, 4]. Example works of modeling behavior include studies of interactions in couples therapy [5, 6, 7], therapist-patient interactions in the addiction domain [8], and behavior in children with autism [9, 10]. For behavior signal processing using lexical information, Georgiou *et al.* [5] used language models to predict the behavior state of a speaker based on spoken language. This used a static behavior model where a persons behavioral state was assumed to remain the same throughout the interaction. A dynamic behavior model to capture transitions between different behavioral states was proposed by Chakravarthula *et al.* [7].

The framework for obtaining (i) *continuous metrics* of behavior, or behavior trajectories, in an (ii) *online*, sliding window manner is crucial to providing psychologists with real-time feedback on patient interactions. This framework should be (iii) robust to *Out-Of-Vocabulary* (OOV) phrases and allow for (iv) long and variable-length *context* where that is available and helpful. Another requirement is to create a framework that easily allows the (v) *fusion* of behavior trajectories with other modalities, such as acoustic features [11]. The following proposed framework as we will demonstrate, meets all our requirements.

Recurrent neural networks (RNN), namely the Long Short-Term Memory (LSTM) architecture, have demonstrated incredible abilities in handling long range dependencies for improved sequence learning [12]. In Natural Language Processing, LSTMs have produced state-of-the-art results in many tasks [13, 14]. However, the use of RNNs in behavior estimation has seen limited success due to data limitations. Firstly, due to privacy restrictions, data with rich information of behavior in psychotherapy sessions is often severely limited in quantity. Secondly, due to the effort required for annotation, very often only the session labels are given and there is no ground truth for individual frames.

In this paper we address these problems and propose an LSTM-RNN system for capturing behavior trajectories in couples interactions in a low data resource environment. To allow for training of the RNN with limited data we use pretrained word representations learned from out-of-domain corpora and joint optimization. We also show the viability of using session-level labels for learning frame-level behavior. Using a fusion of the frame-level behavior trajectories we show that the ratings predicted by our proposed system achieve inter-annotator agreement comparable to those of trained human annotators.

The remaining sections of this paper are organized as

follows: The details of behavior modeling using LSTMs are described in Section 2. Section 3 describes the database used in our training and evaluations. The methodology of our RNN system is detailed in Section 4. Our experimental results and evaluation of the RNN system are presented in Section 5. Finally we present our conclusions in Section 6.

2. Behavior Modeling

2.1. Maximum Likelihood Model

The *Maximum Likelihood* (ML) model assumes that all the utterances observed in a particular session have been generated from the same behavioral state. In previous works [5, 7], we implemented the ML model with n-gram statistical language models of the interlocutor's language. While n-gram language models provide a compact approximation of the joint probability of n-length word sequences, they have limitations. First, the framework suffers when presented with *Out-Of-Vocabulary* (OOV) test data. Secondly ML models are inflexible to variable length n-grams based on data availability (backoff helps but doesn't solve the problem) and this reduces robustness when longer context is introduced. Finally and very importantly, ML models are applicable for classification tasks but not estimation of continuous rating values.

2.2. Behavior Modeling with LSTM

Recurrent neural networks have become increasingly popular for sequence learning tasks as they are adept at integrating temporal information from the entire sequence history as opposed to a fixed window of data in feed-forward neural networks. This dynamic context is especially valuable in natural language processing where semantic meaning may have long-term dependencies across any number of words. RNNs have been shown to perform better than statistical language models in such data-sparse situations by learning distributed representations for words [15, 16]. However the training of RNNs generally requires large amounts of data with accurate labels; something generally not available in our domain. Therefore, we propose the use of pretrained distributed representations of words from out-of-domain large corpora to alleviate the problem of data sparsity. In addition we train the RNN using a weakly supervised method to account for the missing frame-level labels. The details of our proposed RNN system are described in Section 4.

3. Data and Associated Challenges

For our experiments, we use the corpus of 134 couples from the UCLA/UW Couple Therapy Research Project [17]. The dataset contains audio and video recordings, along with transcripts, of real couples with marital issues interacting. In each session, the couples discuss a specific topic (e.g. "why can't you leave my stuff alone ?") chosen in turn for around 10 minutes. The behaviors of each speaker are rated by multiple annotators based on the Couples Interaction [1] and Social Support [18] Rating Systems. This results in 33 behavioral codes such as "Acceptance", "Blame", and "Positivity". Each annotator provides session-level subjective ratings on a Likert scale of 1-9, where 1 indicates absence of the behavior and 9 implies a strong presence. The sessions are rated by 2-12 annotators with majority of the sessions ($\sim$ 90%) rated by 3-4 annotators. Finally, these ratings are averaged to obtain a 33-dimensional vector of session level behavior ratings per interlocutor per interaction.

In this paper, we focus primarily on the "Negativity" behavioral code. As was also done in our earlier work [19, 7, 5] we only consider sessions with mean annotator ratings in the top 20% ('High Negativity') and bottom 20% ('Low Negativity') of the code range for the sessions with good audio quality. This is less than 25% of the whole data set.

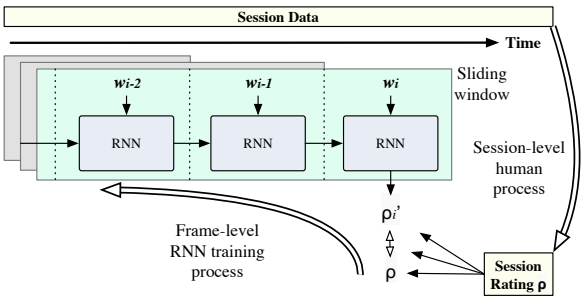

Figure 1: *Training frame-level RNN using global rating values.*

For more information, the reader can refer to [1, 18, 6]

3.1. Associated Challenges

Since human raters do not provide behavioral ratings for each utterance in the session we instead use the global rating as training labels for the individual sequences. In other words, all word sequences within a session are trained with the same label as the global rating. This method assumes that sequences of words from a session are related to global rating in a non-linear, complex manner. This is depicted in Figure 1 where the session-level label ρ is assumed to be a proxy for the frame-label ρ'. This also infers that the longer our context-window the less the mismatch between the global rating ρ and ρ'. Ideally one would like the whole session to be passed as a training sample, however this would drastically decrease our training set and make training difficult. Nevertheless, a larger window can help identify lexical combinations that contribute towards the expression, and consequently estimation, of specific behaviors.

4. Methodology

4.1. Proposed Architecture

We propose a 3-layer recursive neural network architecture as shown in Figure 2.

We encode the input as a one-hot vector $\mathbf{w}$, where the n-th word in our dictionary is represented by setting the n-th element in $\mathbf{w}$ to 1 and all other elements to 0. We assume a vocabulary of N unique words and $0 \leq n \leq N$. The first layer in our

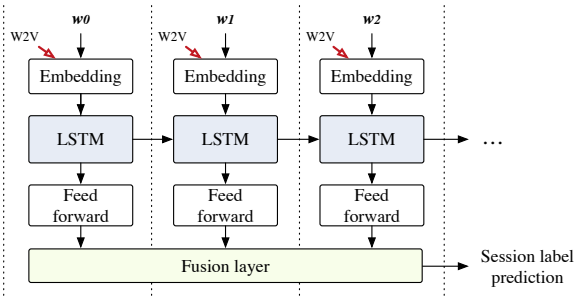

Figure 2: *Recurrent Neural Network system for predicting behavior.*

RNN maps the one-hot vectors $\mathbf{w}$ into intermediate continuous vectors using an embedding layer [20].

The next hidden layer consists of the LSTM blocks that, employing memory cells, will store non-linear representations of the current sequence history and be better able to encode context. To prevent overfitting a dropout layer is added after the LSTM.

Finally the last layer is a feed-forward layer that performs non-linear mappings to better approximate the human scale of behavior. The RNN is then trained for a fixed number of epochs using an adaptive learning rate optimizer [21].

For evaluation purposes, and to better approximate the human annotation process we also require a fusion layer after the RNN to combine the behavior metrics over all the time-steps and obtain a prediction of the global rating.

4.2. Incorporating Out-of-Domain Word Representations

Past work has shown that distributed representations of words in a vector space can be trained to capture syntactic and semantic relationships between words [20, 22]. Such learned representations of words allow learning algorithms to combine semantic knowledge of words and achieve better performance in natural language processing tasks.

In our work, we investigate two options for generating such representations. One is to directly train this on our limited, but domain-specific training data. We will denote this as `1Hot`. Another option that also addresses the problem of data sparsity and allows for a more generalized model, is to incorporate out-of-domain knowledge by pretraining word representations on larger corpora, and we will denote this as `w2v`.

We expect that employing this second method will have advantages: First, by using pre-trained word representations we can mitigate the issue of data sparsity in our training data. High-quality word representations will map similar words to closely spaced points in the vector representation space. This allows us to use a smaller number of parameters and hyper-parameters in constructing and training our RNN. Second, by training on the word representations the system will generalize well in regards to out-of-vocabulary words. Words that were not seen during training will still be mapped to a continuous vector that preserves its semantic relationships to words that were seen during training. The RNN will therefore be able to produce reasonable if not accurate predictions when encountering out-of-vocabulary words in a sequence.

To learn high-quality word representations we use the Google News corpus [23] which contains more than 4 billion words. We also introduce 1 million words from the General Psychotherapy corpus transcripts from [24] to allow the word representations to be more representative of our target domain. The word representations are learned through the methods described in [20] using the Google *word2vec* toolkit [25]. Since our final objective is to estimate the behavior metrics for word sequences we reduce the vector dimensionality from the commonly-used size of 300. In our experiments we tried vector dimensionality configurations of 300, 50, and 10.

The continuous word representations are incorporated into the RNN system by fixing the weights in the embedding layer with the learned word to vector mappings. These weights are then maintained during training to preserve the learned word representations.

4.3. Joint Optimization

Using pretrained word representations the RNN learns to predict the behavior ratings from continuous vectors that capture the semantic relationships between words. However, although these word vectors encode a lot of semantic information they are not optimized for predicting behavior. By jointly training these word vectors with the behavior ratings the word representations become more indicative of behavior where appropriate while still maintaining semantic relationship. In training our RNN with pretrained word representations we initialize with the above learned word vectors and allow the weights in the embedding layer to be updated to allow for this joint optimization. We will denote this by `w2v-joint`.

4.4. Fusion Layer

Our RNN system is trained to predict behavioral ratings for different sequences of words. Since we do not have local-level annotations to compare these predictions with, we evaluate the system at the global session-score level. We do this by fusing the local predictions to arrive at a global predicted score, similar to the human process of integrating behavioral information over time to arrive at a gestalt opinion of the session.

We observed that, in general, the median predicted rating exhibited lesser bias as an estimator of the true rating than the mean rating, possibly due to the former's robustness to outliers. Therefore, we used an RBF-Kernel Support Vector Regressor to learn a mapping from the median predicted rating to the true rating on our training data. At test time, we applied this map on the median predicted test rating to obtain the predicted session-level rating, which we then compared against the true session-level rating that had been used to train our RNN system.

5. Experimental Results

In our experiments we used a leave-one-couple-out cross-validation scheme to separate train and test data. In each fold one couple is held out while the rest are used for training. We applied a sliding window with a 1-word shift across each session to generate multiple training sequences and trained each RNN architecture for 25 epochs. We also tried different dimension sizes for the pretrained word vectors and found that the best results can be obtained from a dimension size of 10.

5.1. Binary Classification of Behavior

We first focused on binary classification of "Negativity" at the session level which is easier to compare with human annotations. A threshold was applied to the average of behavior metrics in a session to classify that session into High or Low Negativity. For each configuration an Equal-Error Rate threshold for the binarization task was obtained from the training data. We trained using different context length for each of the proposed RNN configurations.

The classification accuracy for the different RNN configurations with varying input sequence lengths is shown in Table 1. We observe, as expected due to limited data, a slightly decreasing accuracy as context is increased, but we also see that the accuracy drop is minimal. We also observe that the pretrained word representations (`w2v`) are more robust than

Table 1: *Classification accuracy (%) on negativity for different input sequence lengths.*

RNN Configuration	Input sequence length (words)		
	unigram	bigram	trigram
`1Hot`	87.86	85.71	86.43
`w2v`	87.5	87.1	86.8
`w2v-joint`	88.93	88.21	87.86

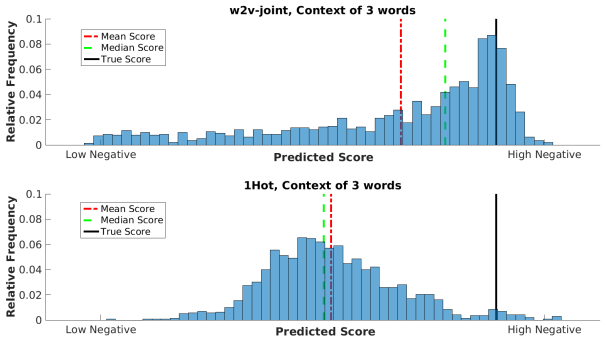

Figure 3: *Comparison of distribution of scores for words in a session for 1-hot system at the bottom versus w2v-joint above.*

embeddings that only employ only domain data (`1hot`) but can become even more robust by joint training (`w2v-joint`).

Note that while the relative improvement is significant it is also limited by the upper limit – even humans do not agree 100% – so the binary evaluation task is limiting our evaluation abilities. For instance, if the upper limit was 100% then we have about 15% relative improvement but if the upper limit is 92% then this jumps to a relative 40% improvement.

5.2. Predicting True Behavior Ratings

5.2.1. Behavioral Distribution

Observing the individual markers of negativity throughout a couples interaction per session we see that the `w2v-joint` system provides a more reasonable distribution of behavior metrics: the behavioral histogram is more skewed towards the true rating value, while the `1-hot` system has very few discriminating data points. For example, Figure 3 shows the distribution of the sequence scores for one session.

5.2.2. Out-Of-Vocabulary Handling

We also analyzed the performance of our RNN system on unseen data: words that were out-of-vocabulary during training. The pretrained system (`w2v-joint`) is able to exploit information from domain-OOV words through their similarity in the general pretraining corpus to seen domain words. Table 2 shows some examples of domain-OOV words and their respective estimated behavior metrics.

Table 2: *Examples of out-of-vocabulary words and their behavior metrics.*

OOV Word	Behavior Metric for Negativity
denies	0.91
kill	0.87
dissatisfaction	0.75
funner	0.26
doggie	0.22
coordination	0.09

5.2.3. Agreement with Human Annotators

To better evaluate our system performance we estimated the behavior ratings which are obtained through the fusion layer. We compared the estimated behavior ratings to those from human annotators using Krippendorff's alpha. In the first comparison we randomly replaced a human annotation with our predicted rating for all sessions. We found that the jointly optimized word representations gave ratings that had better agreement with human ratings than conventional one-hot vectors. Next, we replaced human annotations that deviated most from the mean with our predicted ratings. In this setting we found that our predicted ratings had higher inter-annotator agreement than human-only annotations. This shows that with jointly optimized word representations our RNN system can achieve better inter-annotator agreement than outlier human annotators. The inter-annotator agreement of our predicted ratings for the different comparisons is shown in Table 3.

Table 3: *Comparison of agreement using Krippendorff's alpha.*

Annotator Configuration	Krippendorff's alpha	
	`1Hot`	`w2v-joint`
All human annotators	0.821	
Random replacement with random predictions (average)	0.492	
Random replacement with machine predictions (average)	0.7611	0.7739
Outlier replaced with machine prediction	0.7997	**0.8249**

6. Conclusions

In psychological evaluations of therapy sessions, ratings for behaviors are very often annotated at the global session-level. This coarse resolution drastically increases the difficulty of learning frame-level or utterance-level behaviors. In this paper we have developed a RNN system for estimating behavior in variable-length context windows at the frame level. This enables us to observe continuous metrics of behavior in a sliding window and allows for fusion of behavior from different modalities. The RNN was trained in a data limited environment and only global ratings. We showed that by pretraining word representations on out-of-domain large vocabulary corpora and performing joint optimization we can solve the issue of data sparsity in our data and achieve increased robustness to out-of-vocabulary words. Finally we applied top level fusion on the frame-level behavior metrics to evaluate the behavior trajectories and estimate the true session rating. The estimated behavior rating from our system achieves high agreement with trained human annotators and even outperforms outlier human annotations.

In our work we showed that a RNN system can be trained in a data limited environment to obtain meaningful behavior trajectories in a couples interaction session. This is the first step in allowing for detailed online analysis by psychologists of the interplay of behaviors in couples interactions at a finer resolution. In future work we plan to apply transfer learning between different behavior codes to obtain a better model of complex behaviors. We also plan to build a more complete model through the fusion of behavior metrics from different modalities. For future experiments we plan to also include the noisy unused portions of our data. Current observational studies in psychology often involve the time-consuming and expensive process of annotating specific behaviors in lengthy sessions. In the future we hope to deploy our system for a more automated method of evaluating behavior in human interactions.

7. References

[1] C. Heavey, D. Gill, and A. Christensen, "Couples interaction rating system 2 (cirs2)," *University of California, Los Angeles*, vol. 7, 2002.

[2] G. Margolin, P. H. Oliver, E. B. Gordis, H. G. O'hearn, A. M. Medina, C. M. Ghosh, and L. Morland, "The nuts and bolts of behavioral observation of marital and family interaction," *Clinical child and family psychology review*, vol. 1, no. 4, pp. 195–213, 12 1998.

[3] P. G. Georgiou, M. P. Black, and S. S. Narayanan, "Behavioral signal processing for understanding (distressed) dyadic interactions: Some recent developments," in *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*. ACM, 2011, pp. 7–12.

[4] 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.

[5] P. G. Georgiou, M. P. Black, A. Lammert, B. Baucom, and S. S. Narayanan, ""That's aggravating, very aggravating": Is it possible to classify behaviors in couple interactions using automatically derived lexical features?" in *Proceedings of Affective Computing and Intelligent Interaction*, Memphis, TN, USA, 2011.

[6] M. Black, A. Katsamanis, B. Baucom, C. Lee, A. Lammert, A. Christensen, P. Georgiou, and S. Narayanan, "Toward automating a human behavioral coding system for married couples interactions using speech acoustic features," *Speech Communication*, 2012.

[7] S. N. Chakravarthula, R. Gupta, B. Baucom, and P. Georgiou, "A language-based generative model framework for behavioral analysis of couples' therapy," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, Apr. 2015.

[8] B. Xiao, D. Bone, M. Van Segbroeck, Z. E. Imel, D. Atkins, P. Georgiou, and S. Narayanan, "Modeling therapist empathy through prosody in drug addiction counseling," in *Proceedings of Interspeech*, Sep. 2014.

[9] E. Mower, C.-C. Lee, J. Gibson, T. Chaspari, M. Williams, and S. S. Narayanan, "Analyzing the nature of eca interactions in children with autism," in *Proceedings of Interspeech*, Aug. 2011.

[10] T. Chaspari, C.-C. Lee, and S. S. Narayanan, "Interplay between verbal response latency and physiology of children with autism during eca interactions," in *Proceedings of InterSpeech*, Sep. 2012.

[11] W. Xia, J. Gibson, B. Xiao, B. Baucom, and P. Georgiou, "A dynamic model for behavioral analysis of couple interactions using acoustic features," in *Proceedings of Interspeech*, September 2015.

[12] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[13] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling." in *Proceedings of Interspeech*, 2012, pp. 194–197.

[14] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005.

[15] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain, *Innovations in Machine Learning: Theory and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, ch. Neural Probabilistic Language Models, pp. 137–186.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.

[17] A. Christensen, D. Atkins, S. Berns, J. Wheeler, D. Baucom, and L. Simpson, "Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples," *Journal of Consulting and Clinical Psychology*, vol. 72, no. 2, pp. 176–191, 2004.

[18] J. Jones and A. Christensen, "Couples interaction study: Social support interaction rating system," University of California, Los Angeles, Technical manual, 1998.

[19] M. P. Black, A. Katsamanis, C.-C. Lee, A. Lammert, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Automatic classification of married couples' behavior using audio features," in *Proceedings of InterSpeech*, 2010.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *In Proceedings of Workshop at ICLR*, 2013.

[21] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *The Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2011.

[22] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations." in *HLT-NAACL*, 2013, pp. 746–751.

[23] "Google news corpus." [Online]. Available: http://www.statmt.org/wmt14/training-monolingual-news-crawl/

[24] "General psychotherapy corpus." in *Alexander Street Press*. [Online]. Available: http://alexanderstreet.com

[25] "Google word2vec toolkit." [Online]. Available: http://word2vec.googlecode.com