Grounding Gradable Adjectives through Crowdsourcing

Rebecca Sharp, Mithun Paul, Ajay Nagesh, Dane Bell, and Mihai Surdeanu
CLU Lab, Dept. of Computer Science, University of Arizona
{bsharp, mithunpaul, ajaynagesh, dane, msurdeanu}@email.arizona.edu

Abstract

In order to build technology that has the ability to answer questions relevant to national and global security, e.g., on food insecurity in certain parts of the world, one has to implement machine reading technology that extracts causal mechanisms from texts. Unfortunately, many of these texts describe these interactions using vague, high-level language. One particular example is the use of gradable adjectives, i.e., adjectives that can take a range of magnitudes such as *small* or *slight*. Here we propose a method for estimating specific concrete groundings for a set of such gradable adjectives. We use crowdsourcing to gather human language intuitions about the impact of each adjective, then fit a linear mixed effects model to this data. The resulting model is able to estimate the impact of novel instances of these adjectives found in text. We evaluate our model in terms of its ability to generalize to unseen data and find that it has a predictive R^2 of 0.632 in general, and 0.677 on a subset of high-frequency adjectives.

1. Intro

In order to understand the interplay of various entities and events in complex systems such as climate change or crop yields, scientists make use of complex quantitative models (e.g., DSSAT (Jones et al., 2003) or AGMIP (Rosenzweig et al., 2013)) with qualitative hypotheses (e.g., Parry and Rosenzweig (1993), Zhang el al. (2011), Zhao et al. (2017), *inter alia*). Typically these models must be hand-generated by domain experts through an expensive process that requires extensive literature review and a large amount of time, resulting in only a small fraction of the available information being processed and incorporated into the models. These models are crucial to predicting the vagaries in such complex systems. A timely and accurate assessment of the factors that affect such systems bears directly on the health of the nation and environment (Elliott et al., 2017; Porwollik et al., 2017). Automated machine reading can help create and validate hypotheses using these models, but there remains a disconnect – often relevant events are written in high-level language, yet the model requires specific quantities. In particular, when describing events, authors often make use of *gradable adjectives*, i.e., adjectives (such as *small*) that can take a range of magnitudes or degrees (e.g., something can be a little small, very small, extraordinarily small, etc.)

When used to describe important changes in model parameters, as with this snippet from a scientific publication: *"...doubling of the atmospheric carbon dioxide concentration will lead to only a* ***small decrease*** *in global crop production."* (Parry and Rosenzweig, 1993), the ability to quantify or ground such adjectives is a critical step for automated machine reading. For example, consider the difference between a *small increase in rainfall* versus a *severe increase in rainfall* when attempting to predict the funds needed for disaster relief or the potential impact on the expected yield of a given crop.

Here we propose a method for quickly and concretely grounding a large number of gradable adjectives such that their effect can be calculated for entities or events of interest. Specifically, we gather human intuitions about the effect of a particular gradable adjective on a given distribution (i.e., mean and standard deviation), independent of the item being modified, and then use this data to fit a linear model. Gradable adjectives are often classified in terms of scales, e.g., *hot, warm*, and *cool* are all in the temperature scale while *large* and *small* describe magnitude instead. Here we focus on these magnitude adjectives as a use case, though we suggest that the method can be straightforwardly extended to other scales.

The resulting resource we provide[1] consists of a linear model for each adjective that takes as input the typical distribution of the item being modified and in turn provides the predicted size of the change. Turning once again to the example above, for a region with a average rainfall of 40 inches/year (±6 in), if we need to ground a *small* increase, the model would return a predicted rainfall of 40.54 inches. This can be compared, for example, to 45.26, which is the model's prediction for a *large* increase.

Our specific contributions are:

- We provide a method for using human language intuitions about the semantic meaning of gradable adjectives to create a viable model for the adjective semantics. We decouple the meaning of the adjective from the noun being modified to get a truer semantic understanding of the adjective itself, and we show that the model predictions are well correlated with the human judgments.
- Using cross validation, i.e., testing on crowd-sourced predictions not seen in training but whose adjectives were seen in training, we show that we can achieve a predictive R^2 of 0.632 when using all adjectives. When we use a smaller subset of higher frequency adjectives, we can fit a higher precision model that has a predictive R^2 of 0.677. One limitation of this model is that it is unable to make predictions on novel, or unseen, adjectives. We address this with an initial neural network model based on word embeddings. On unseen adjectives this model has a predictive R^2 of 0.244.
- We release the resulting database of 98 adjectives

[1] All materials, data, and code used are available at `https://github.com/clulab/releases/tree/master/lrec2018-gradable`

and their corresponding linear models as a domain-agnostic resource. This resource could potentially facilitate the quantification of extracted entities and events that contain gradable adjectives for use in downstream tasks, such as modeling complex systems and predicting real-world events. To address a variety of future use cases, we break this resource down into several versions: the full version with all adjectives that emphasizes recall, a smaller subset that emphasizes precision, and a second version of each of these that depends only on the adjective (i.e., for when the typical distribution of the noun being modified is unknown).

2. Related Work

Gradable adjectives have been experimentally demonstrated to be interpreted in part based on the semantics of the nouns they modify (i.e., a *small* mouse versus a *small* building) (Bonini et al., 1999; Alxatib and Pelletier, 2011; Bylinina, 2014). For this reason, we test our adjectives using legal non-words (e.g. *mards*), for which we provide a typical size distribution, and we include the provided distribution in the linear model. This allows us to model the semantics of the adjective in context, while removing the need to test each adjective along with each possible noun it could modify.

Whitman et al. (2003) propose a model for finding perceptual groundings for adjectives. That is, they ground their adjectives using audio features and are then able to predict a lexical description of unheard music. This is quite similar to what we do in spirit, however our groundings are numerical rather than perceptual. Several works have attempted to link gradable adjectives with numerical quantities that co-occur in the context of the gradable adjective mention (Shivade et al., 2016; Narisawa et al., 2013). However, this dependence on corpus resources to find evidence for gradability requires complex information extraction and suffers greatly from sparsity, especially when attempting to ground adjectives in a new domain. Additionally, it results in a solution that is highly domain-specific (Shivade et al., 2016).

Kim and de Marneffe repurposed neural network language models by using word embeddings to rank gradable adjectives (2013). Bakhshandeh and Allen (2015) use bootstrapping to discover properties of adjectives, including what they can modify. However, rather than simply ranking adjectives or extracting their attributes, here we focus on determining a concrete, numerical grounding for each.

There have also been recent works that use crowdsourcing to determine gradability of adjectives. For example, Qing and Franke (2014) use crowdsourcing to gather intuitions about the interpretation of gradable adjectives, but they are testing whether or not adjective usage corresponds to optimal language use, whereas our resource grounds gradable adjectives. Accordingly, they test only four adjectives using visual cues, while we test 98 adjectives using numerical cues.

The work by Wilkinson and Tim (2016) is closest to our work. They use crowdsourcing to create ranked lists of gradable adjectives that correspond to a variety of different scales (e.g., temperature, dimension, speed, and so on). Unlike Wilkinson and Tim, we use only one scale (magnitude) and again, we are interested in creating a concrete grounding for adjectives rather than a ranking.

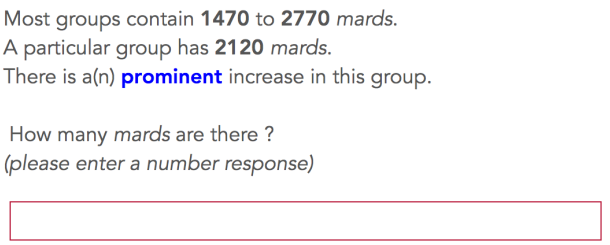

Figure 1: Example prompt given to Amazon Mechanical Turk workers to elicit the impact of gradable adjectives. Workers were given a specific distribution (of an imaginary item) and asked for the increase they perceived from the given adjective. The full set of prompts is included with our release.

3. Approach

While humans often use high-level language to describe events, models of the interactions between these events require specific quantitative information. To bridge the gap between gradable adjectives and this quantified representation, we use human language intuitions about an adjective's impact on a given distribution to fit a linear model. With this model, then, we can predict the impact of the adjective on an entity whose distribution is known.

Specifically our approach operates in two parts: gathering the human language intuitions for each adjective using crowdsourcing (described in Section 4.) and then fitting a linear mixed effects model to the data (described in Section 5.). The resulting model allows us to make predictions about the effects of one of our grounded adjectives on unseen nouns.

4. Data

We first gathered a set of 98 gradable adjectives from the Collins Birmingham University International Language Database (COBUILD) dictionary (Sinclair and others, 1987) that were determined to be particularly relevant to use cases focusing on national and global security (e.g., food insecurity) and able to be evaluated with a common methodology. We then used Amazon Mechanical Turk (MTurk) (Buhrmester et al., 2011) to gather our data for each of these adjectives. Our MTurk task was designed to test the amount by which a given adjective was perceived to change a known quantity. As discussed in Section 2., the perceived impact of a gradable adjective has been shown to be highly dependent on the typical distribution of the item it is modifying (Bonini et al., 1999; Alxatib and Pelletier, 2011; Bylinina, 2014). For example, an extra 2 inches of rain would be insignificant in a tropical location, but in the desert it would be a large change. However, since we want a model that can be used in a range of contexts, we designed our experiment to decouple the adjective semantics from the noun semantics by using non-words for the items

being modified (in the style of a "Wug Test" (Berko, 1958)) and providing MTurk workers (turkers) with a typical distribution of the item in question, as shown in the example prompt in Figure 1. We then ask the turkers to describe the effect of the adjective in question on the group size. This response forms the basis for the dependent variable used in our model building.

We required the turkers to be in the United States and informed them that they needed to be native speakers of English. To demonstrate this, they were required to correctly answer a language-based question in order to participate. They were given two attempts (i.e., a second question was shown to them if they did not correctly answer the first). For the task, each turker was asked to provide responses to 16 prompts and was compensated with $0.75, based on the average of 20 seconds per prompt.

5. Model Building

5.1. Model Factors

As interpretations of gradable adjectives are context-dependent (Section 2.), in addition to the factor of interest (i.e., adjective) we include in our model the shape of the distribution for the item being modified using two control factors: the provided mean (μ_p) and provided standard deviation (σ_p). We calculate σ_p directly from the typical range (e.g., 1470 to 2770 in the example in Figure 1), which we consider to be ± 2 standard deviations.[2] Thus, for the above example, $\sigma_p = (2770 - 1470)/4 = 325$. The value for the particular group given to turkers (i.e., 2120) lies in the middle of the range and so we use that directly for μ_p[3]. We chose not to add the interactions between these factors and adjective due to the large number of degrees of freedom (recall that adjective has 98 levels) in an effort to reduce the likelihood of overfitting.

In addition to differences based on context, gradable adjectives have also been shown to be interpreted differently by different individuals (Raffman, 1994; Raffman, 1996; Shapiro, 2006). To account for this, we elected to fit a linear mixed effects model to our data, as this allows us to include a random intercept for each turker. In effect, this means that while we are fitting a linear model with adjective, σ_p, and μ_p as fixed effects, we allow the fitted line to have a different intercept for each turker, thereby accounting for individual biases.[4] While it is possible in this framework to also have random slopes for each adjective, here we refrained so as to avoid a large increase in model complexity.

[2]The interpretation of *most* in terms of standard deviations is ambiguous, and may vary between domains. While we have chosen to build our model under the interpretation of *most* as ± 2 standard deviations, the resulting model could be calibrated to a specific domain by gathering a small set of adjective instances that are accompanied by a specific value. We leave this to future work.

[3]In a pilot study we found that neither the direction of the change (i.e., *increase* vs. *decrease*) nor the non-word used significantly affected the model, so in this study we did not include these as factors.

[4]Note that while random intercepts allow the model to be more robust to variance due to individual biases, the model included in our final resource are averaged across respondents, thus allowing predictions for novel instances.

Fixed Effects	χ^2	p-value
μ_p:σ_p	$\chi^2\,(1) = 1.98$	$p = 0.16$
μ_p	$\chi^2\,(1) = 5.59$	$p < 0.05$
σ_p	$\chi^2\,(97) = 151.46$	$p < 0.001$

Table 1: Results of the likelihood ratio tests (LRTs) used to determine the significance of the model's fixed effects. Significance was determined through a likelihood ratio test comparing a model with the predictor to a model without.

The dependent variable (i.e., what the model is trying to predict) is the response given by the turkers, normalized by μ_p and σ_p into something very similar to a z-score:

$$respDev = \frac{|response - \mu_p|}{\sigma_p} \quad (1)$$

In this way, a $respDev$ of 0.5 indicates an increase of 0.5 standard deviations from the mean. Boxplots showing the responses for a subset of the adjectives are shown in Figure 2. The collected values for these adjectives align with human intuitions, and we also see something of a floor effect whereby the responses for the adjectives that indicate a smaller change seem to have much lower variance than the responses for the adjectives that indicate a larger change. For example, Figure 2 highlights that there is a small variance in responses for *conservative* and *slight*, but a the large variance in responses for *huge* and *major*.

5.2. Data Cleaning

We initially gathered 50 data points for each of our 98 adjectives. However, responses generated using MTurk can be quite noisy, so to reduce the amount of noise in our data we excluded data based on several criteria. We excluded all responses from turkers we considered to be unreliable because more than 50% of their responses were outliers[5], 20% or more of their responses were identical to μ_p, or 50% or more of their responses were identical to one of the given range endpoints. We also removed responses that were less than or equal to the mean, as all prompts asked for an amount of increase. We then removed outliers by adjective. Finally, we removed responses from turkers who had 4 or fewer responses remaining (for the purposes of model fitting and evaluation). This left a total of 3309 responses for our 98 adjectives.

5.3. Model Fitting

Our model fitting was done using the `lme4` package in R (Bates et al., 2015; R Core Team, 2013). The residuals of the model (i.e., $respDev \sim 1 + adj + \sigma_p + \mu_p + \sigma_p : \mu_p + (1|turker)$) showed that the data was heteroskedastic. That is, as the predicted values from the model increased, so did the magnitudes of the error residuals. To adjust for this, we log-transformed $respDev$ to create $logRespDev$. After verifying that the resulting residual plot showed homoskedasticity (see Figure 3), we used $logRespDev$ in all subsequent model-building.

[5]We considered any points more than 1.5 times the interquartile range below the first quartile or above the third quartile to be outliers.

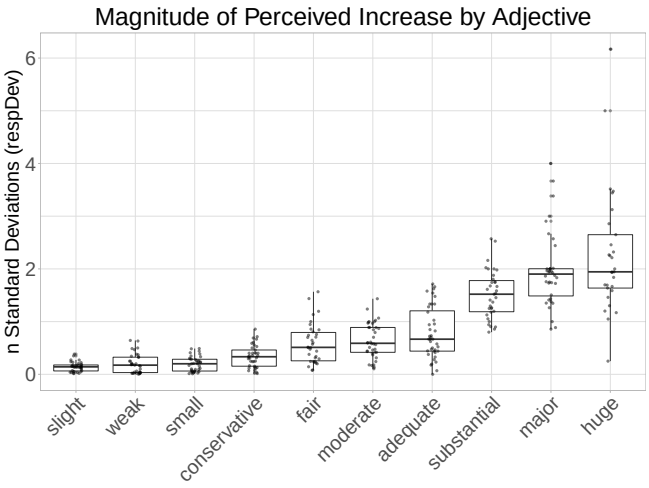

Figure 2: The magnitudes of the perceived increase for several adjectives. The magnitudes are measured as the absolute difference between the survey response and the given mean divided by the given standard deviation.

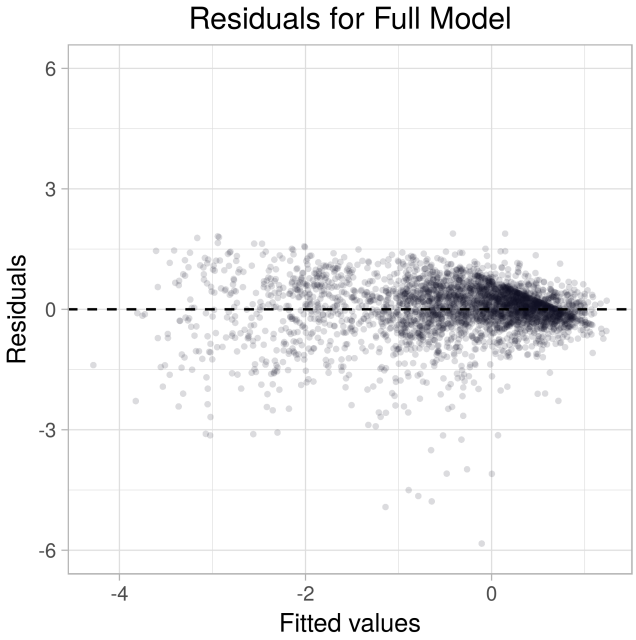

Figure 3: Residual plot for the model with all factors included: $logRespDev \sim 1 + adj + \sigma_p + \mu_p + \sigma_p : \mu_p + (1|turker)$. The residual plot for the final model (with $\sigma_p : \mu_p$ removed) is omitted for space, but it is nearly identical.

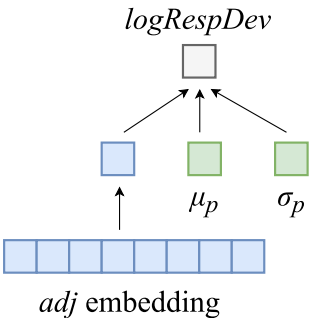

Figure 4: Shallow neural network architecture for predicting the impact of adjectives that were not seen during training using the word embedding of the adjective. As with the linear models, we include the provided mean (μ_p) and provided standard deviation (σ_p) as factors and predict the log-transformed response deviations ($logRespDev$).

We used likelihood ratio tests (LRTs) to determine the significance of the fixed effects (i.e., our three factors and their interactions) by first building a parent model that included all factors as well a daughter models that each had a factor removed. We then checked to see if the model with the factor removed was significantly different from its parent model. The resulting significances are shown in Table 1. As σ_p:μ_p was first determined to not be significant, it was removed and the model without this factor was used as the parent model for testing the significance of σ_p and μ_p, both of which were determined to be significant (see Table 1). The final model is given by:

$$logRespDev \sim 1 + adj + \sigma_p + \mu_p + (1|turker) \quad (2)$$

This fitted model itself is our resource. That is, for each adjective, we have a linear function, $f_{adj}(\mu_p, \sigma_p)$ that describes its predicted impact of the quantity in question. For example, the adjective *small* is represented as:

$$f_{small}(\mu_p, \sigma_p) = -1.77 + (1.034\text{e}{-5})\mu_p - (1.123\text{e}{-3})\sigma_p$$

The predicted new value implied by *small* can then be calculated from this as:

$$new = (e^{f_{small}(\mu_p,\sigma_p)} \times \sigma_p) + \mu_p$$

6. Alternative Models

6.1. Backoff Model

Though the standard deviation was significant, it is not always the case that this will be known. For this situation, we also provide a backoff model that does not include σ_p. The dependent variable used in this model is the absolute percent change in the mean (log-transformed):

$$logPercChange \sim 1 + adj + \mu_p + (1|turker) \quad (3)$$

6.2. High Frequency (HF) Models

Under the hypothesis that language intuitions will better align for more commonly used words, we additionally retrained our fitted model and backoff models on a smaller subset of the data that consists of the highest frequency words We sorted the adjectives based on their frequency in the English Gigaword corpus (Graff et al., 2003) and retained only the top 30 adjectives. We used this higher frequency subset of adjectives to train a regular as well as a backoff model.

6.3. Grounding Using Neural Networks

Even though the number of adjectives covered by these linear models is fairly large (98), they are unable to predict groundings for novel, or *unseen*, adjectives. To address this limitation, we propose a shallow neural network (NN) model which builds upon pre-trained word embeddings (Mikolov et al., 2013). Intuitively, by using word embeddings trained over a large corpus, we already know some of the underlying semantics of the unseen adjectives.

	Full	Backoff	HF	HF-Backoff
marginal R^2	0.618	0.548	0.672	0.562
conditional R^2	0.670	0.589	0.725	0.596
predictive R^2	0.632	0.544	0.677	0.542

Table 2: Estimation from all models of how much of the variance in the data is accounted for by the model's fixed effects (marginal R^2), and both the fixed and random effects (conditional R^2). Also, an measure of how well the model predicts new data, predictive R^2.

Therefore, to the extent that the embedding of a given adjective captures its implied magnitude, we can learn a mapping from this embedding to the specific, quantitative grounding for the adjective.

In this NN approach, shown in Figure 4, we use a fully-connected hidden layer of size one to compress the adjective's high-dimensional word embedding to a single value (shown in blue), the activation of which can be directly interpreted as the semantic impact of the adjective learned by the model. This value is then concatenated to the provided mean (μ_p) and provided standard deviation (σ_p) and passed to an output layer that predicts the the log transformed response deviation ($logRespDev$). In this framework we found that the features that uniquely identify the individual respondent from the crowd-sourcing experiment were not needed (i.e., they did not greatly improve performance), and so we removed them to help prevent overfitting.

We trained the model using mean squared error as the loss function. The embeddings were initialized with the Glove (Pennington et al., 2014) 300-dimensional pre-trained word embeddings and they were not updated during training (again, to reduce overfitting). A tanh activation was used for the adjective node[6] and we use the RMSProp optimizer (Tieleman and Hinton, 2012) with a learning rate of 0.00001 and all other parameters with their default values.

7. Results

7.1. Linear Models

For the evaluation of our linear models, we report the marginal and conditional R^2 in Table 2. The marginal R^2 shows the amount of the variance that is explained by only the fixed effects and the conditional R^2 shows the amount of the variance that is explained by both the fixed and random effects. Both were calculated using the `r.squaredGLMM` function from the MuMIn (Barton, 2016) package, an implementation of the method of Nakagawa and Schielzeth (2013). As we are primarily interested in using this resource to make predictions about new instances of adjectives, the correlation of the model's predictions with real data is key. Thus, we also calculate the predictive R^2 with leave-one-out cross-validation, such that the residual error of each data point (i.e., individual response) is based on a model trained on all the data except for that point. Specifically, the predictive R^2 is the predicted residual error sum of squares (PRESS) statistic

[6]We tried using non-linear activations on all the nodes but did no see an improvement so we omitted them for model simplicity.

	Seen Adjs	Unseen Adjs
Linear Full Model	0.632	–
NN Model	0.540	0.244

Table 3: Comparison of how well the linear mixed effects full model and the neural network (NN) model predict new data (predictive R^2). Performance is shown for predictions both on adjectives that were present in the training data (seen) as well as on adjectives that were not (unseen).

(Allen, 1974) divided by the total sum of squares (SS_{total}):

$$R^2_{pred} = 1 - \frac{PRESS}{SS_{total}} \quad (4)$$

$$PRESS = \sum_{i=1}^{n} (logRespDev_i - \widehat{logRespDev}_{i,D\setminus i})^2 \quad (5)$$

$$SS_{total} = \sum_{i=1}^{n} (logRespDev_i - \overline{logRespDev})^2 \quad (6)$$

That is, for each individual response $i \in D$, we sum the residual squared error between the true value, $logRespDev_i$ and the value predicted by a model trained on the rest of the data, $\widehat{logRespDev}_{i,D\setminus i}$. We then divide this by SS_{total} and subtract if from 1 to get the predictive R^2. For our full model, the predictive R^2 was 0.632 (also shown in Table 2). This result suggests that the quantities implied by these adjectives can be predicted with reasonable accuracy with simple, linear models trained on crowd-sourced data.

We found that the backoff subset model had a slightly worse fit than the full model. This is expected as it does not contain the standard deviation as a factor, which was determined to be significant (Section 5.3.).

The high-frequency (HF) model shows a higher R^2 than the full model. This confirms that, indeed, language intuitions are more robust for high-frequency adjectives. However, this effect is only seen in the full model when standard deviation is known.

7.2. Neural Network Model

We evaluate our neural network (NN) model on both seen and unseen adjectives. The predictive R^2 on *seen* adjectives (i.e., when data points for each adjective are split between training and test folds) can be compared to the performance of the linear models, while the performance on *unseen* adjectives (i.e., when adjectives appearing in test folds do *not* occur in training folds) indicates utility with novel adjectives. Due to time constraints rather than using leave-one-out cross-validation (as with the linear models) we instead use four-fold cross-validation with two folds for training, one for development, and one for testing.

The number of epochs for each fold was tuned on this split to avoid overfitting. The other hyperparameters (e.g., learning rate) were tuned on 10% of the data in early experiments and were not revisited. The resulting predictive R^2s are shown in Table 3.

On seen adjectives, the NN model performs almost as well as the linear model, and we suspect that the performance difference is primarily due to the larger number of parameters that need to be learned. Additionally, while the information about individual turkers was empirically found to

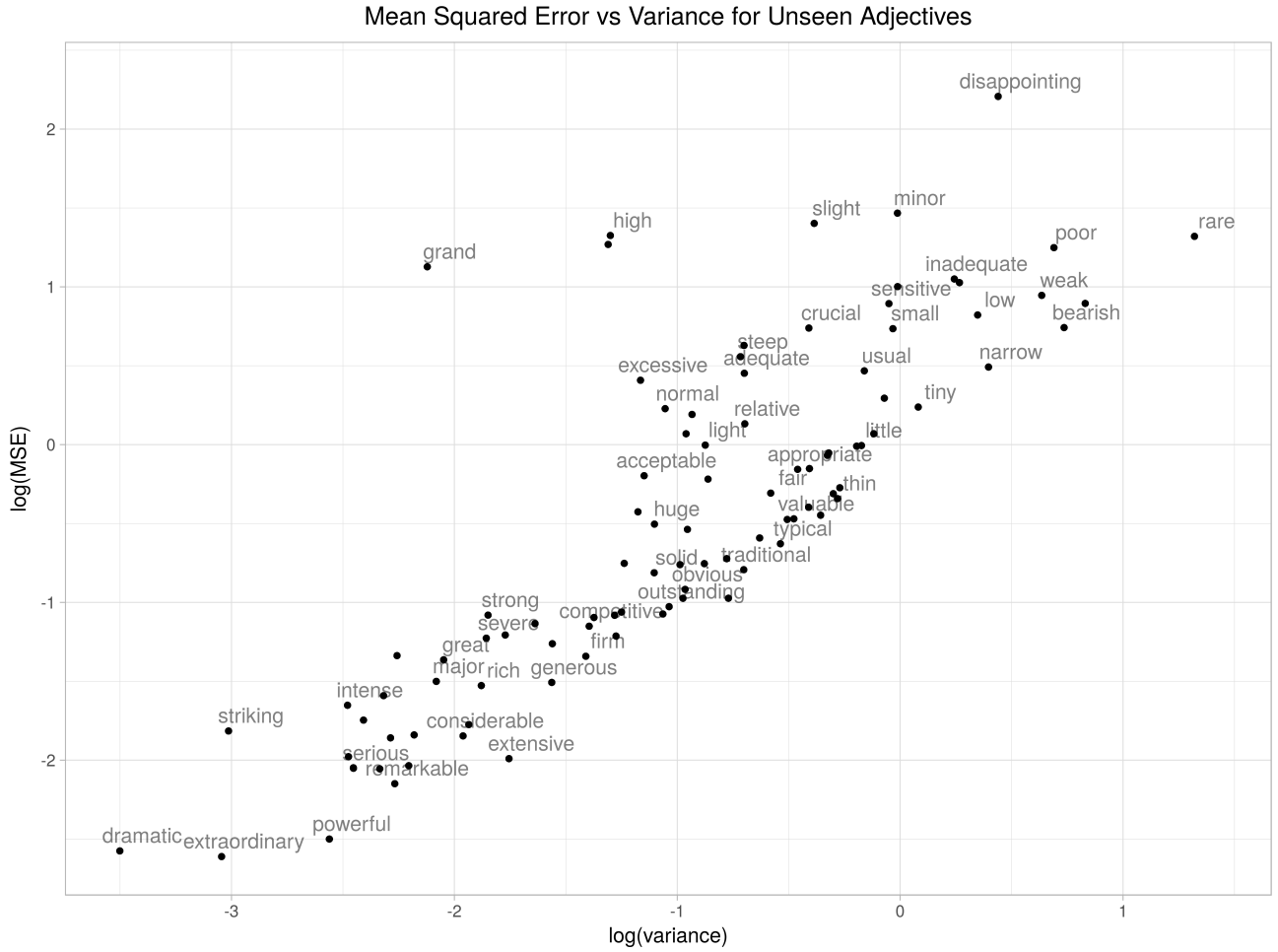

Figure 5: A comparison of the mean squared error (MSE) versus variance for each gradable adjective. The MSE is based on the neural network model predicting on unseen adjectives and the variance is from the original crowd-sourced data. Note that the axes are presented in log scale.

not help the NN model, the linear model benefits from its inclusion.

The predictive R^2 for unseen adjectives is much lower than for seen adjectives, but recall that the linear model is unable to make any predictions for these adjectives at all. It is unclear exactly where the performance drop originates, though we hypothesize that it is primarily due to the reliance on pre-trained word-embeddings. While they allow us to estimate groundings for adjectives that we did not include in training, the estimates are only as good as our capacity to extract the necessary information from the embedding. That is, the embeddings were trained to capture distributional similarity, not relative magnitude of impact. Thus, this information, when present, is indirect and very likely noisy.

To better understand the performance of this model, we compared the mean squared error (MSE) of the model on these unseen adjectives with their variance in the original data from the crowd-sourcing experiment. The plot is shown in Figure 5. Overall, as the variance in the original data increases, so does the MSE. This suggests that in general adjectives with higher variance are harder to predict. Further, some adjectives had much higher variance (e.g., *rare* and *disappointing*), suggesting that for some adjectives this task is difficult even for humans. Additionally, certain adjectives (such as *grand, high*, and *disappointing*) had a particularly high MSE. We suspect that this is, again, due to the reliance on the pre-trained emebeddings as many of these words have multiple senses (e.g., *disappointing*, as with *disappointing increase* versus *the play was disappointing*), and here the sense we are interested in is not the most frequent. For these words with multiple senses, the embeddings are confounded. To address these issues, dedicated embeddings that better model the semantics of interest could be explored (such as the embeddings proposed by Kim et al. (2016) that are dedicated for modeling adjectives). We leave this exploration to future work.

8. Conclusion

We proposed a method for quickly and efficiently generating groundings for a set of gradable adjectives. These groundings are modeled using a linear model conditioned on the typical distribution of the item being modified while remaining otherwise independent of the item's identity. The model was trained on approximately 50 values collected through crowdsourcing for each of the adjectives in the set. The resulting model has a predictive R^2 of 0.632 on the whole dataset (measured through leave-one-

out cross-validation), and a R^2 of 0.677 on a subset of high-frequency adjectives. We release all models created for these adjectives, which, we hope, brings us closer to developing technology that answers questions relevant to national and global security from texts containing qualitative statements.

9. Acknowledgements

This work was funded by the Defense Advanced Research Projects Agency (DARPA) World Modeling Seed program under ARO contract W911NF-17-1-0047.

10. Bibliographical References

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16(1):125–127.

Alxatib, S. and Pelletier, F. J. (2011). The psychology of vagueness: Borderline cases and contradictions. *Mind & Language*, 26(3):287–326.

Bakhshandeh, O. and Allen, J. (2015). From adjective glosses to attribute concepts: Learning different aspects that an adjective can describe. In *IWCS*.

Barton, K. (2016). Mumin: Multi-model inference (r package version 1.15. 6.)[computer software].

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Berko, J. (1958). The child's learning of English morphology. *Word*, 14(2-3):150–177.

Bonini, N., Osherson, D., Viale, R., and Williamson, T. (1999). On the psychology of vague predicates. *Mind & language*, 14(4):377–393.

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's mechanical turk. *Perspectives on Psychological Science*, 6(1):3–5.

Bylinina, E. (2014). *The Grammar of Standards. Judge-dependence, Purpose-relativity, and Comparison Classes in Degree Constructions*. Utrecht University.

Elliott, J., Glotter, M., Ruane, A. C., Boote, K. J., Hatfield, J. L., Jones, J. W., Rosenzweig, C., Smith, L. A., and Foster, I. (2017). Characterizing agricultural impacts of recent large-scale us droughts and changing technology and management. *Agricultural Systems*.

Graff, D., Kong, J., Chen, K., and Maeda, K. (2003). English gigaword, ldc2003t05. *Linguistic Data Consortium, Philadelphia*.

Jones, J., Hoogenboom, G., Porter, C., Boote, K., Batchelor, W., Hunt, L., Wilkens, P., Singh, U., Gijsman, A., and Ritchie, J. (2003). The dssat cropping system model. *European Journal of Agronomy*, 18(3):235–265. Modelling Cropping Systems: Science, Software and Applications.

Kim, J. and de Marneffe, M. (2013). Deriving adjectival scales from continuous space word representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1625–1630.

Kim, J.-K., de Marneffe, M.-C., and Fosler-Lussier, E. (2016). Adjusting word embeddings with semantic intensity orders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 62–69.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Nakagawa, S. and Schielzeth, H. (2013). A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142.

Narisawa, K., Watanabe, Y., Mizuno, J., Okazaki, N., and Inui, K. (2013). Is a 204 cm man tall or small? acquisition of numerical common sense from the web. In *ACL (1)*, pages 382–391. The Association for Computer Linguistics.

Parry, M. and Rosenzweig, C., (1993). *The Potential Effects of Climate Change on World Food Supply*, pages 1–26. Springer Berlin Heidelberg, Berlin, Heidelberg.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Porwollik, V., Müller, C., Elliott, J., Chryssanthacopoulos, J., Iizumi, T., Ray, D. K., Ruane, A. C., Arneth, A., Balkovič, J., Ciais, P., Deryng, D., Folberth, C., Izaurralde, R. C., Jones, C. D., Khabarov, N., Lawrence, P. J., Liu, W., Pugh, T. A., Reddy, A., Sakurai, G., Schmid, E., Wang, X., de Wit, A., and Wu, X. (2017). Spatial and temporal uncertainty of crop yield aggregations. *European Journal of Agronomy*, 88(Supplement C):10–21. Uncertainty in crop model predictions.

Qing, C. and Franke, M. (2014). Meaning and use of gradable adjectives: Formal modeling meets empirical data. In Paul Bello, et al., editors, *Proceedings of the 36th annual meeting of the Cognitive Science Society (CogSci-2014)*, pages 1204–1209, Austin, TX. Cognitive Science Society.

R Core Team, (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raffman, D. (1994). Vagueness without paradox. *The Philosophical Review*, 103(1):41–74.

Raffman, D. (1996). Vagueness and context-relativity. *Philosophical Studies*, 81(2-3):175–192.

Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., Boote, K. J., Thorburn, P., Antle, J. M., Nelson, G. C., Porter, C., Janssen, S., et al. (2013). The agricultural model intercomparison and improvement project (agmip): protocols and pilot studies. *Agricultural and Forest Meteorology*, 170:166–182.

Shapiro, S. (2006). *Vagueness in context*. Oxford University Press on Demand.

Shivade, C., de Marneffe, M., Fosler-Lussier, E., and Lai, A. M. (2016). Identification, characterization, and grounding of gradable terms in clinical text. In *Proceedings of the 15th Workshop on Biomedical Natural*

Language Processing, BioNLP@ACL 2016, Berlin, Germany, August 12, 2016, pages 17–26.

Sinclair, J. et al. (1987). *Collins COBUILD English language dictionary*. Harper Collins Publishers,.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning.

Whitman, B., Roy, D., and Vercoe, B. (2003). Learning word meanings and descriptive parameter spaces from music. In *Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non-linguistic Data - Volume 6*, HLT-NAACL-LWM '04, pages 92–99, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wilkinson, B. and Tim, O. (2016). A gold standard for scalar adjectives. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Zhang, D. D., Lee, H. F., Wang, C., Li, B., Pei, Q., Zhang, J., and An, Y. (2011). The causality analysis of climate change and large-scale human crisis. *Proceedings of the National Academy of Sciences*, 108(42):17296–17301.

Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., Huang, M., Yao, Y., Bassu, S., Ciais, P., Durand, J.-L., Elliott, J., Ewert, F., Janssens, I. A., Li, T., Lin, E., Liu, Q., Martre, P., Müller, C., Peng, S., Peñuelas, J., Ruane, A. C., Wallach, D., Wang, T., Wu, D., Liu, Z., Zhu, Y., Zhu, Z., and Asseng, S. (2017). Temperature increase reduces global yields of major crops in four independent estimates. *Proceedings of the National Academy of Sciences*, 114(35):9326–9331.