Interpreting Neural Networks to Improve Politeness Comprehension

Malika Aubakirova
University of Chicago
aubakirova@uchicago.edu

Mohit Bansal
UNC Chapel Hill
mbansal@cs.unc.edu

Abstract

We present an interpretable neural network approach to predicting and understanding politeness in natural language requests. Our models are based on simple convolutional neural networks directly on raw text, avoiding any manual identification of complex sentiment or syntactic features, while performing better than such feature-based models from previous work. More importantly, we use the challenging task of politeness prediction as a testbed to next present a much-needed understanding of what these successful networks are actually learning. For this, we present several network visualizations based on activation clusters, first derivative saliency, and embedding space transformations, helping us automatically identify several subtle linguistics markers of politeness theories. Further, this analysis reveals multiple novel, high-scoring politeness strategies which, when added back as new features, reduce the accuracy gap between the original featurized system and the neural model, thus providing a clear quantitative interpretation of the success of these neural networks.

1 Introduction

Politeness theories (Brown and Levinson, 1987; Gu, 1990; Bargiela-Chiappini, 2003) include key components such as modality, indirection, deference, and impersonalization. Positive politeness strategies focus on making the hearer feel good through offers, promises, and jokes. Negative politeness examples include favor seeking, orders, and requests. Differentiating among politeness types is a highly nontrivial task, because it depends on factors such as a context, relative power, and culture. Danescu-Niculescu-Mizil et al. (2013) proposed a useful computational framework for predicting politeness in natural language requests by designing various lexical and syntactic features about key politeness theories, e.g., first or second person start vs. plural. However, manually identifying such politeness features is very challenging, because there exist several complex theories and politeness in natural language is often realized via subtle markers and non-literal cues.

Neural networks have been achieving high performance in sentiment analysis tasks, via their ability to automatically learn short and long range spatial relations. However, it is hard to interpret and explain what they have learned. In this paper, we first propose to address politeness prediction via simple CNNs working directly on the raw text. This helps us avoid the need for any complex, manually-defined linguistic features, while still performing better than such featurized systems. More importantly, we next present an intuitive interpretation of what these successful neural networks are learning, using the challenging politeness task as a testbed.

To this end, we present several visualization strategies: activation clustering, first derivative saliency, and embedding space transformations, some of which are inspired by similar strategies in computer vision (Erhan et al., 2009; Simonyan et al., 2014; Girshick et al., 2014), and have also been recently adopted in NLP for recurrent neural networks (Li et al., 2016; Kádár et al., 2016). The neuron activation clustering method not only rediscovers and extends several manually defined features from politeness theories, but also uncovers multiple novel strategies, whose importance we measure quantitatively. The first derivative saliency technique allows us to identify the impact of each phrase

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2035–2041,
Austin, Texas, November 1-5, 2016.

on the final politeness prediction score via heatmaps, revealing useful politeness markers and cues. Finally, we also plot lexical embeddings before and after training, showing how specific politeness markers move and cluster based on their polarity. Such visualization strategies should also be useful for understanding similar state-of-the-art neural network models on various other NLP tasks.

Importantly, our activation clusters reveal two novel politeness strategies, namely indefinite pronouns and punctuation. Both strategies display high politeness and top-quartile scores (as defined by Danescu-Niculescu-Mizil et al. (2013)). Also, when added back as new features to the original featurized system, they improve its performance and reduce the accuracy gap between the featurized system and the neural model, thus providing a clear, quantitative interpretation of the success of these neural networks in automatically learning useful features.

2 Related Work

Danescu-Niculescu-Mizil et al. (2013) presented one of the first useful datasets and computational approaches to politeness theories (Brown and Levinson, 1987; Goldsmith, 2007; Kádár and Haugh, 2013; Locher and Watts, 2005), using manually defined lexical and syntactic features. Substantial previous work has employed machine learning models for other sentiment analysis style tasks (Pang et al., 2002; Pang and Lee, 2004; Kennedy and Inkpen, 2006; Go et al., 2009; Ghiassi et al., 2013). Recent work has also applied neural network based models to sentiment analysis tasks (Chen et al., 2011; Socher et al., 2013; Moraes et al., 2013; Dong et al., 2014; dos Santos and Gatti, 2014; Kalchbrenner et al., 2014). However, none of the above methods focused on visualizing and understanding the inner workings of these successful neural networks.

There have been a number of visualization techniques explored for neural networks in computer vision (Krizhevsky et al., 2012; Simonyan et al., 2014; Zeiler and Fergus, 2014; Samek et al., 2016; Mahendran and Vedaldi, 2015). Recently in NLP, Li et al. (2016) successfully adopt computer vision techniques, namely first-order *saliency*, and present representation plotting for sentiment compositionality across RNN variants. Similarly, Kádár et al. (2016) analyze the omission scores and top-k contexts of hidden units of a multimodal RNN. Karpathy et al. (2016) visualize character-level language models. We instead adopt visualization techniques for CNN style models for NLP[1] and apply these to the challenging task of politeness prediction, which often involves identifying subtle and non-literal sociolinguistic cues. We also present a quantitative interpretation of the success of these CNNs on the politeness prediction task, based on closing the performance gap between the featurized and neural models.

3 Approach

3.1 Convolutional Neural Networks

We use one convolutional layer followed by a pooling layer. For a sentence $v_{1:n}$ (where each word v_i is a d-dim vector), a filter m applied on a window of t words, produces a convolution feature $c_i = f(m * v_{i:i+t-1} + b)$, where f is a non-linear function, and b is a bias term. A *feature map* $c \in R^{n-t+1}$ is applied on each possible window of words so that $c = [c_1, ..., c_{n-t+1}]$. This convolutional layer is then followed by a max-over-pooling operation (Collobert et al., 2011) that gives $C = max\{c\}$ of the particular filter. To obtain multiple features, we use multiple filters of varying window sizes. The result is then passed to a fully-connected softmax layer that outputs probabilities over labels.

4 Experimental Setup

4.1 Datasets

We used the two datasets released by Danescu-Niculescu-Mizil et al. (2013): Wikipedia (Wiki) and Stack Exchange (SE), containing community requests with politeness labels. Their 'feature development' was done on the Wiki dataset, and SE was used as the 'feature transfer' domain. We use a simpler train-validation-test split based setup for these datasets instead of the original leave-one-out cross-validation setup, which makes training extremely slow for any neural network or sizable classifier.[2]

[1] The same techniques can also be applied to RNN models.

[2] The result trends and visualizations using cross-validation were similar to our current results, in preliminary experiments. We will release our exact dataset split details.

4.2 Training Details

Our tuned hyperparameters values (on the dev set of Wiki) are a mini-batch size of 32, a learning rate of 0.001 for the Adam (Kingma and Ba, 2015) optimizer, a dropout rate of 0.5, CNN filter windows of 3, 4, and 5 with 75 feature maps each, and ReLU as the non-linear function (Nair and Hinton, 2010). For convolution layers, we use valid padding and strides of all ones. We followed Danescu-Niculescu-Mizil et al. (2013) in using SE only as a transfer domain, i.e., we do not re-tune any hyperparameters or features on this domain and simply use the chosen values from the Wiki setting. The split and other training details are provided in the supplement.

5 Results

Table 1 first presents our reproduced classification accuracy test results (two labels: positive or negative politeness) for the bag-of-words and linguistic features based models of Danescu-Niculescu-Mizil et al. (2013) (for our dataset splits) as well as the performance of our CNN model. As seen, without using any manually defined, theory-inspired linguistic features, the simple CNN model performs better than the feature-based methods.[3]

Next, we also show how the linguistic features baseline improves on adding our novelly discovered features (plus correcting some exising features), revealed via the analysis in Sec. 6. Thus, this reduces the gap in performance between the linguistic features baseline and the CNN, and in turn provides a quantitative reasoning for the success of the CNN model. More details in Sec. 6.

6 Analysis and Visualization

We present the primary interest and contribution of this work: performing an important qualitative and quantitative analysis of what is being learned by our neural networks w.r.t. politeness strategies.[4]

6.1 Activation Clusters

Activation clustering is a non-parametric approach (adopted from Girshick et al. (2014)) of computing each CNN unit's activations on a dataset and then analyzing the top-scoring samples in each cluster. We keep track of which neurons get maximally activated for which Wikipedia requests and analyze the most frequent requests in each neuron's cluster, to understand what each neuron reacts to.

Model	Wiki	SE
Bag-of-Words	80.9%	64.6%
Linguistic Features	82.6%	65.2%
With Discovered Features	83.8%	65.7%
CNN	85.8%	66.4%

Table 1: Accuracy Results on Wikipedia and Stack Exchange.

6.1.1 Rediscovering Existing Strategies

We find that the different activation clusters of our neural network automatically rediscover a number of strategies from politeness theories considered in Danescu-Niculescu-Mizil et al. (2013) (see Table 3 in their paper). We present a few such strategies here with their supporting examples, and the rest (e.g., Gratitude, Greeting, Positive Lexicon, and Counterfactual Modal) are presented in the supplement. The majority politeness label of each category is indicated by (+) and (-).

Deference (+) A way of sharing the burden of a request placed on the addressee. Activation cluster examples: {"*nice work so far on your rewrite...*"; "*hey, good work on the new pages...*"}

Direct Question (-) Questions imposed on the converser in a direct manner with a demand of a factual answer. Activation cluster examples: {"*what's with the radio , and fist in the air?*"; "*what level warning is appropriate?*"}

6.1.2 Extending Existing Strategies

We also found that certain activation clusters depicted interesting extensions of the politeness strategies given in previous work.

Gratitude (+) Our CNN learns a special shade of gratitude, namely it distinguishes a cluster consisting of the bigram *thanks for*. Activation cluster examples: {"*thanks for the good advice.*"; "*thanks for letting me know.*"}

Counterfactual Modal (+) Sentences with *Would you/Could you* get grouped together as expected; but in addition, the cluster contains requests with *Do you mind* as well as gapped 3-grams like *Can you ... please?*, which presumably implies that the combi-

[3]For reference, human performance on the original task setup of Danescu-Niculescu-Mizil et al. (2013) was 86.72% and 80.89% on the Wiki and SE datasets, respectively.

[4]We only use the Wiki train/dev sets for all analysis.

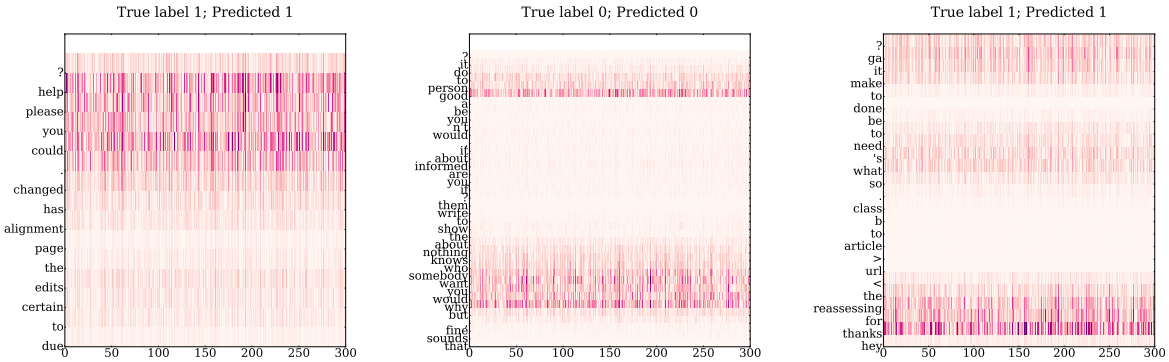

Figure 1: Saliency heatmaps for correctly classified sentences.

nation of a later *please* with future-oriented variants *can/will* in the request gives a similar effect as the conditional-oriented variants *would/could*. Activation cluster examples: {*can this be reported ... grid, please?*"; *do you mind having another look?*"}

6.1.3 Discovering Novel Strategies

In addition to rediscovering and extending politeness strategies mentioned in previous work, our network also automatically discovers some novel activation clusters, potentially corresponding to new politeness strategies.

Indefinite Pronouns (-) Danescu-Niculescu-Mizil et al. (2013) distinguishes requests with first and second person (plural, starting position, etc.). However, we find activations that also react to indefinite pronouns such as *something/somebody*. Activation cluster examples: {"*am i missing something here?*"; "*wait for anyone to discuss it.*"}

Punctuation (-) Though non-characteristic in direct speech, punctuation appears to be an important special marker in online communities, which in some sense captures verbal emotion in text. E.g., one of our neuron clusters gets activated on question marks "???" and one on ellipsis "...". Activation cluster examples: {"*now???*"; "*original article????*"; "*helllo?????*"}[5]

In the next section, via saliency heatmaps, we will further study the impact of indefinite pronouns in the final-decision making of the classifier. Finally, in Sec. 6.4, we will quantitatively show how our newly discovered strategies help directly improve the accuracy performance of the linguistic features baseline and achieve high politeness and top-quartile scores as per Danescu-Niculescu-Mizil et al. (2013).

[5]More examples are given in the supplement.

6.2 First Derivative Saliency

Inspired from neural network visualization in computer vision (Simonyan et al., 2014), the first derivative saliency method indicates how much each input unit contributes to the final decision of the classifier. If E is the input embedding, y is the true label, and $S_y(E)$ is the neural network output, then we consider gradients $\frac{\partial S_y(E)}{\partial e}$. Each image in Fig. 1 is a heatmap of the magnitudes of the derivative in absolute value with respect to each dimension.

The first heatmap gets signals from *please* (Please strategy) and *could you* (Counterfactual Modal strategy), but effectively puts much more mass on *help*. This is presumably due to the nature of Wikipedia requests such that the meaning boils down to asking for some help that reduces the social distance. In the second figure, the highest emphasis is put on *why would you*, conceivably used by Wikipedia administrators as an indicator of questioning. Also, the indefinite pronoun *somebody* makes a relatively high impact on the decision. This relates back to the activation clustering mentioned in the previous section, where indefinite pronouns had their own cluster. In the third heatmap, the neural network does not put much weight on the greeting-based start *hey*, because it instead focuses on the higher polarity[6] gratitude part after the greeting, i.e., on the words *thanks for*. This will be further connected in Sec. 6.3.

6.3 Embedding Space Transformations

We selected key words from Danescu-Niculescu-Mizil et al. (2013) and from our new activation clusters (Sec. 6.1) and plotted (via PCA) their embed-

[6]See Table 3 of Danescu-Niculescu-Mizil et al. (2013) for polarity scores of the various strategies.

	Strategy	Politeness	In top quartile	Examples
21.	Indefinite Pronouns	-0.13	39%	*am i missing something here?*
22.	Punctuation	-0.71	62%	*helllo?????*

Table 2: Extending Table 3 of Danescu-Niculescu-Mizil et al. (2013) with our novelly discovered politeness strategies.

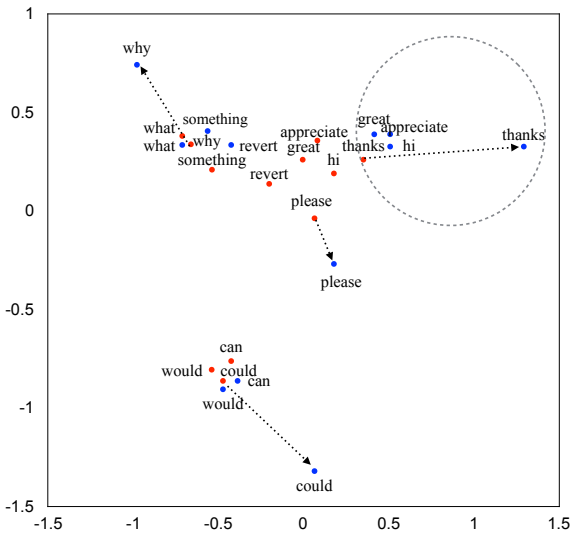

Figure 2: Projection before (red) and after (blue) training.

ding space positions before and after training, to help us gain insights into specific sentiment transformations. Fig. 2 shows that the most positive keys such as *hi*, *appreciate*, and *great* get clustered even more tightly after training. The key *thanks* gets a notably separated position on a positive spectrum, signifying its importance in the NN's decision-making (also depicted via the saliency heatmaps in Sec. 6.2).

The indefinite pronoun *something* is located near direct question politeness strategy keys *why* and *what*. *Please*, as was shown by Danescu-Niculescu-Mizil et al. (2013), is not always a positive word because its sentiment depends on its sentence position, and it moves further away from a positive key group. Counterfactual Modal keys *could* and *would* as well as *can* of indicative modal get far more separated from positive keys. Moreover, after the training, the distance between *could* and *would* increases but it gets preserved between *can* and *would*, which might suggest that *could* has a far stronger sentiment.

6.4 Quantitative Analysis

In this section, we present quantitative measures of the importance and polarity of the novelly discovered politeness strategies in the above sections, as well how they explain some of the improved performance of the neural model.

In Table 3 of Danescu-Niculescu-Mizil et al. (2013), the pronoun politeness strategy with the highest percentage in top quartile is 2nd Person (30%). Our extension Table 2 shows that our novelly discovered Indefinite Pronouns strategy represents a higher percentage (39%), with a politeness score of -0.13. Moreover, our Punctuation strategy also turns out to be a top scoring negative politeness strategy and in the top three among all strategies (after Gratitude and Deference). It has a score of -0.71, whereas the second top negative politeness strategy (Direct Start) has a much lower score of -0.43.

Finally, in terms of accuracies, our newly discovered features of Indefinite Pronouns and Punctuation improved the featurized system of Danescu-Niculescu-Mizil et al. (2013) (see Table 1).[7] This reduction of performance gap w.r.t. the CNN partially explains the success of these neural models in automatically learning useful linguistic features.

7 Conclusion

We presented an interpretable neural network approach to politeness prediction. Our simple CNN model improves over previous work with manually-defined features. More importantly, we then understand the reasons for these improvements via three visualization techniques and discover some novel high-scoring politeness strategies which, in turn, quantitatively explain part of the performance gap between the featurized and neural models.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. This work was supported by an IBM Faculty Award, a Bloomberg Research Grant, and an NVIDIA GPU donation to MB.

[7] Our NN visualizations also led to an interesting feature correction. In the 'With Discovered Features' result in Table 1, we also removed the existing pronoun features (#14-18) based on the observation that those had weaker activation and saliency contributions (and lower top-quartile %) than the new indefinite pronoun feature. This correction and adding the two new features contributed ∼50-50 to the total accuracy improvement.

References

Francesca Bargiela-Chiappini. 2003. Face and politeness: new (insights) for old (concepts). *Journal of pragmatics*, 35(10):1453–1469.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Long-Sheng Chen, Cheng-Hsiang Liu, and Hui-Ju Chiu. 2011. A neural network based approach for sentiment classification in the blogosphere. *Journal of Informetrics*, 5(2):313–322.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of ACL*.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of ACL*, pages 49–54.

Cícero Nogueira dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING*, pages 69–78.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2009. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341.

M Ghiassi, J Skinner, and D Zimbra. 2013. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications*, 40(16):6266–6282.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of CVPR*, pages 580–587.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1:12.

Daena J Goldsmith. 2007. Brown and levinsons politeness theory. *Explaining communication: Contemporary theories and exemplars*, pages 219–236.

Yueguo Gu. 1990. Politeness phenomena in modern chinese. *Journal of pragmatics*, 14(2):237–257.

Dániel Z Kádár and Michael Haugh. 2013. *Understanding politeness*. Cambridge University Press.

Akos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2016. Representation of linguistic form and function in recurrent neural networks. *arXiv preprint arXiv:1602.08952*.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of ACL*.

Andrej Karpathy, Justin Johnson, and Fei-Fei Li. 2016. Visualizing and understanding recurrent networks. In *Proceedings of ICLR Workshop*.

Alistair Kennedy and Diana Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2):110–125.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of NIPS*, pages 1097–1105.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of NAACL*.

Miriam A Locher and Richard J Watts. 2005. Politeness theory and relational work. *Journal of Politeness Research. Language, Behaviour, Culture*, 1(1):9–33.

Aravindh Mahendran and Andrea Vedaldi. 2015. Understanding deep image representations by inverting them. In *Proceedings of CVPR*, pages 5188–5196. IEEE.

Rodrigo Moraes, Joao Francisco Valiati, and Wilson P GaviãO Neto. 2013. Document-level sentiment classification: An empirical comparison between svm and ann. *Expert Systems with Applications*, 40(2):621–633.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of ICML*, pages 807–814.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL*, page 271.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Bach, and Klaus-Robert Müller. 2016. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of ICLR Workshop*.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Proceedings of ECCV*, pages 818–833. Springer.