How do we get there? Evaluating transformer neural networks as cognitive models for English past tense inflection

Anonymous ACL submission

Abstract

Neural network models have achieved good performance on morphological inflection tasks, including English past tense inflection. However whether they can represent human cognitive mechanisms is still under debate. In this work, we examined transformer models with different size and distribution of training data to show that: 1) neural model's performance correlates with the adult behavior, but not children's behavior; and the model with small-size training data that matches parents' input distribution has the highest correlation; 2) neural models' errors are not human-like; however, the errors on the regulars and irregulars show a clear distinction. Therefore, we conclude that the current transformer models exhibit some resemblance of human behavior, but is insufficient as a cognitive model of learning morphological rules.

1 Introduction

English past tense has been the subject of debate in human language processing for decades. The past tense has attracted so much attention because both adults and children exhibit a clear distinction between the regulars and irregulars. The regular form follows a formal rule: adding '-ed [*/d/,/t/,/ɪd/*]' to the verb stem as in 'help/helped' applies to a vast majority of English verbs and can be generalized to novel words by adults and children (e.g. 'wug-wugged', Berko (1958)). The irregular forms consist of $\sim$ 200 verbs in English. Some of the patterns can be categorized by phonological similarities, as in 'sing/sang', 'sink/sank', 'drink/drank and 'begin/began', but these patterns are rarely generalized by human speakers. Thus, the debate of English past tense has been focused on the nature of the regular-irregular distinction, whether it is a *discrete* distinction that is governed by rules (e.g. Pinker and Prince, 1988), or a *gradient* distinction that is generated by phonological analogy (e.g. Bybee and Moder, 1983).

Rumelhart and McClelland (1986) (hence RM) first proposed that past tense inflection can be learned by the neural model. They constructed a connectionist model that learns to associate phonological features of the stem with phonological features of the past-tense forms. Since the early fixed-size feed-forward network can't handle sequences with varied lengths, they constructed wickelfeatures based on wickephones (Wickelgren, 1969) as input. Each wickelfeature is a phonological feature set of a trigram in the root verb, e.g. /ɛlp/ is represented as [<+vowel, +continuous, +unvoiced> + <+low, +liquid, +stop>]. The model successfully learned the regular and irregular forms. RM's model received fierce criticisms that the neural network is susceptible to the frequency distribution: it may learn the *most frequent* pattern, but not the *regular* pattern (Marcus et al., 1992).

Recent works on encoder-decoder (ED) neural networks as cognitive models have focused on the generalization ability, namely, does the neural network have human-like performance on nonce verb production. For English past tense, Kirov and Cotterell (2018) showed that the ED RNN model is able to generalized '-ed' to nonce verbs like adult speakers; however Corkery et al. (2019) showed that the correlation between the ED RNN model's performance and adult speakers' production is weak.

This study further investigates this issue by comparing the model's performance to both adults' and children's production. In particular, we ask the following questions: 1) Does the neural network model's performance correlate with human adults and/or children's behavior in nonce verb production? 2) Are the errors on real verbs child-like? What are the characteristics of the errors? In this work, instead of using RNN, we use the transformer model since it is the state-of-art system for langauge modeling in NLP. We begin by showing that transformer models with different training sizes all

significantly correlate with human adult's data, but only the model with children's input distribution correlates with the children's data. The model's errors are not child-like, but exhibit a distinction between regulars and irregulars. We conclude that the transformer model shows some resemblance of human behavior, but is insufficient as a cognitive model of morphological rule learning.

2 Background

2.1 Nonce verb experiment

With adult participants. One of the most replicated nonce verb experiments is Albright and Hayes (2003) (hence AH). They created a set of 58 nonce verbs that are similar to the English verbs. Each nonce verb was assigned a regular past tense form, e.g. 'gleed /glid/' - 'gleeded /glid**ɪd**/' and an irregular past form, e.g. 'gled /gl**ɛd**/'. The participants were asked to produce the past tense of each nonce verb, as well as rate the regular and irregular forms. Each form's **production probability** and **ratings** were calculated. In general, the human participants predominately produced the regular form for most of the nonce verbs.

In addition, they also constructed a rule-based model and an analogy model which predicts an acceptance score for the past tense forms. The analogy model's score is calculated based on the phonological similarity[1] of each nonce verb to the existing verbs in the CELEX (Baayen et al., 1995) database of English verbs (4253 verbs, 218 of which are irregulars). For example, the score for regular form '/glidɪd/' is calculated based on phonologically similarities to the regular verbs such as 'speed', 'need'; the score for irregular form '/glɛd/' is calculated based on the similarities to the irregular verbs such as 'bleed', 'feed'. The rule-based model's score is calculated based on the proportion of existing verbs that can be explained by certain linguistic rules. For example, the regular form '/glidɪd/' is formed based on the regular rule: '+ /ɪd/' if verb matches [X /d/,/t/__], e.g. 'want', 'need', which explains 87.2% past tense forms of the verbs ending in /d/ or /t/; thus the score for '/glidɪd/' is 0.872. The irregular form '/glɛd/' is generated based on an irregular rule: '/i/>/ɛ/ if verb matches [X /r/,/l/ __/d/], e.g. 'bleed', 'read', which explains 79.3% past tens forms of verbs that matches [X /r/,/l/ __/d/]; thus the score for '/glɛd/' is 0.793. AH compared the analogy model's score and rule-based model's score with human participants' production abilities and rating on each nonce verb's regular and irregular past tense form. They concluded that the rule-based model correlates with human speaker's behavior better than the analogy model.

With children participants. The nonce verb experiment has also been replicated on children. Blything et al. (2018) used the same 40 nonce verbs and recruited children from 4 age groups (3-4 y/o, 5-6 y/o, 6-7 y/o and 9-10 y/o) for a production task. The older children produced more regular forms than the younger children. In addiction, children also produced non past-tense forms of the nonce verb, such as the verb root, 3rd singular ('-s') and progressive ('-ing').

With neural models. Kirov and Cotterell (2018) (hence KC) revisited the past tense debate with a biLSTM encoder-decoder model. They used a subset of verbs in the CELEX dataset, which contains 4039 verbs, 168 of which are irregular. Their model reached near-perfect accuracy for the regular verbs (98.9%) and also achieved some accuracy for irregular verbs (28.6%). They also showed that the encoder-decoder model effectively models human behavior in nonce verbs. The correlation of model's nonce verb output is significantly correlated with human production probabilities (Spearman's ρ = 0.48 for regulars and ρ = 0.45 for irregulars).

Corkery et al. (2019) (hence CMS) also conducted the a similar nonce verb experiments on biLSTM models, but did not find such strong correlations. They adopted the model architecture in KC and trained the model on all 4253 verbs as in AH and 4039 verbs in KC. They used the beam probabilities of each regular and irregular form to calculate the correlation with human data. They showed that with different random initializations, the model's output correlates with the human production probability differently, ranging from ρ = 0.1 - 0.6 for regulars and ρ = 0.2 - 0.4 for irregulars. They wondered if these models should be treated as individual participants instead of an averaged representation. Therefore, they further trained 50 individual models with same training data and hyperparameters and sampled 100 past tense forms from each model to have an aggregated model result. The aggregated model shows better correlations than individual model, but still not as good as the rule-based model in AH. CMS also suspected

[1]The phonological similarity is measured based on the natural class theory by Broe (1993).

that 100 training epochs might lead to model overfitting, and training for less time might have better correlations with human data. Reducing training epochs to 10 achieved the best correlation with human data, but resulted in bad accuracy on real verbs.

2.2 Children's errors on past tense

English speaking children's past tense error has been one of the most widely studied phenomenon in linguistics and psychology. The past tense acquisition has been characterized by **overregularization error**. (e.g. Plunkett and Marchman, 1991; Marcus et al., 1992; Xu and Pinker, 1995; Maratsos, 2000; Maslen et al., 2004). Overregularization errors are the incorrect past forms of irregular verbs when children add '-ed [/d/,/t/,/ɪd/]' to the stem. The most common type of overregularization errors is 'Stem+ed', e.g. '*drawed', '*falled', '*maked'. Children also attach '-ed' to the irregular form ('Past+ed'), such as '*boughted', '*felled', '*tored'. In addition, previous studies also found other rare errors such as incorrect vowel change, e.g. 'bring-*brang' on irregulars.

2.3 Evaluating model

Human Behavior. In this work, we first correlate the model's output on nonce verbs with both **production probability** and **rating** data for adults and children. In addition, we conduct error analysis on the nonce verbs to examine the differences of model's prediction and human production. We also compare the model's errors on real English verbs with children's overregularization errors to see if the model mimics children's errors.

Cognitive theories. We further compare the transformer model with the rule-based model and the analogy model by correlating the model's output on nonce verbs with the **acceptance score** predicted by the two models reported in AH.

3 Methods

3.1 Architecture and hyperparameters

We use the transformer model for our training. The transformer model is a self-attention-based encoder-decoder model that is able to process sequential data in a parallel manner, which is different from the LSTM models. The transformer model has achieved great success in complex tasks like machine translation and language generation. Since the datasets for our character-level morphological inflection task are significantly smaller than traditional transformer tasks, we employed a smaller transformer with 2 layers in the encoder (1 attention layer, 1 feed-forward layer) and 3 layers in the decoder (2 attention layers, 1 feed-forward layer). Layer normalization is applied to the output of encoder and decoder. Positional embedding layers are used to capture the positional information. We use 6 self-attention heads, embedding size is 256 and hidden size of feed-forward layer is 1024. The transformer model has $\sim$ 5.83M parameters. Training was done using Adadelta optimization (Zeiler, 2012) with batch size of 32. We train 100 epochs for each model.

3.2 Models and Data

Modeling Adults. To counter the overfitting problem mentioned in CMS, we decide to reduce the training data instead of reducing the number of epochs. We randomly sampled 500, 1500 and 3000 verbs as training data from 4039 verbs used in KC. We also adopt CMS's idea that each model should be treated as an individual participant. CMS changed the initializations of each model to generate different 'participants'. We change the training data for each model by randomly generating 30 samples with 500 verbs, 1500 verbs and 3000 verbs to create 30 'participants' for each training size. We aggregate 30 participants models' output for each training size to produce the **models' production probability**. In the training data, the average proportion of irregular is 4% for models with 500, 1500 and 3000 verbs.

Modeling Children. Children are exposed to less verbs than adults with higher proportion of irregulars. To better model the verbs that children are exposed to, we generate the training data based on real-life parents' input verbs. We selected 8 children's corpora in the CHILDES database (MacWhinney, 2000) that contain overregularization errors. We included each child's first recording file to the first file where they made overregularization errors, and aggregated the parents' the past tense verbs, which contains 246 unique past tense verbs (65 irregular verbs) [2]. The irregular proportion is 26%, which is higher than other training datasets. We randomly generated 30 samples with 246 verbs in CELEX dataset matching the numbers

[2]The detailed summary of parent's data in shown in Table9 in Appendix.

of regular and irregular verbs in the parents' input as our training set and aggregate these models output to produce production probability. The detailed proportion of regular and irregular verbs in each training set is shown in Table 1.

Data size	Regular % /-d/	/-t/	/-ɪd/	Irregular irr%
500	50 (2.2)	19 (2.2)	27 (0.7)	4 (0.7)
1500	51 (1.2)	18 (0.9)	27 (0.9)	4 (0.4)
3000	51 (0.5)	18 (0.4)	27 (0.4)	4 (0.2)
246	42	22	10	26

Table 1: The mean proportions of regulars and irregulars (standard deviation in brackets) averaged over 30 samples of training data with different size

Test Data. We evaluate the models on the nonce verbs and real English verbs. We use all 58 unique nonce verbs for comparing adult's behavior, matching AH, and 40 nonce verbs matching Blything et al. (2018) to compare children's behavior. We also randomly selected 150 regular verbs (50 for */d/*, */t/* and */ɪd/*) and 20 irregular verbs from the CELEX dataset as the testing data for real English verbs.

4 Experiments

First we report the train accuracy as a sanity check in Table 2. All models achieved almost perfect training accuracy on the regulars and over 90% training accuracy on the irregulars, showing that the model successfully learned the past tense forms during training. The small training size model has the best training accuracy on the irregulars, since this model also has higher proportion of irregulars in the training data.

Data size	Regular %	Irregular %
246	99.45 (0.04)	95.31 (1.29)
500	99.37 (0.03)	90.56 (0.82)
1500	99.86 (0.03)	91.67 (0.78)
3000	99.84 (0.03)	90.83 (0.81)

Table 2: Mean training accuracy (standard deviations in brackets) averaged over 30 samples for each data size.

4.1 Experiment 1: Correlation with human data

4.1.1 Correlation with adults' behavior

We calculated the correlation between the model's production probability and adult's production probability and ratings using Pearson's r. The results are listed in Table 3.

Rating: *Between Regular and Irregular*: All the models are significantly correlated with the adult's rating for both regulars and irregulars. The correlation with regulars are generally higher than the irregulars, but the differences are not significant. *Among models*: The model with 246 verbs has highest correlation with regulars and irregulars. Increasing the training size of the model does not result in higher correlation. Instead, small-training-size model seems to correlate with adult ratings better. Our models correlate with the adult ratings better than CMS and KC. Only the model with 246 verbs matching parents' distribution perform better than the rule-based and analogy model.

Production probability: *Between Regular and Irregular*: All models are significantly correlated with the production probability for regulars. For irregulars, the models with 3000 verbs and 1500 verbs are not significantly correlated with production probability. In general, the correlation for regulars are higher than irregulars, but there is no significant differences. *Among models*: Similar to the rating, the model with 246 verbs has higher correlation. There is no significant differences among correlations. The model with 246 verbs also correlates better than the rule-based model and the analogy model.

Summary: In general, most of our models show significant correlations with production probability and rating for both regulars and irregulars. The models have higher correlations with regulars than irregulars. Model with 246 verbs correlates with adult's production probability and rating better than other models. It is puzzling that models with more training verbs did not have better correlation. One possible explanation is that the irregular proportion in the model with 246 verbs (26%) is higher than other models, which better represents the verbs distribution that adults exposed to.

4.1.2 Correlation with children's behavior

We only used the 3-4 y/o children's data in our study. Only the model with 246 verbs is significantly correlated with irregulars for the children data. No other significant correlations were found.

4.1.3 Correlation with Cognitive Models

Between Regular and Irregular: For regulars, all models are significantly correlated with the rule-based model and the analogy model except for the

Correlation	Regular			Irregular		
	Adult		Children (3-4 y/o)	Adult		Children (3-4 y/o)
r	Production Probability	Rating	Production Probability	Production Probability	Rating	Production Probability
246	**0.67**	**0.77**	0.11	**0.75**	**0.66**	**0.63**
500	**0.47**	**0.53**	0.01	**0.35**	**0.38**	0.14
1500	**0.41**	**0.46**	-0.1	0.21	**0.30**	-0.06
3000	**0.50**	**0.52**	-0.11	0.2	**0.29**	-0.08
Rule-based	**0.62**	**0.70**		**0.31**	**0.46**	
Analogy	**0.56**	**0.59**		**0.13**	**0.45**	
CMS	*0.30*	*0.4*		*0.17*	*0.40*	
KC	*0.48*			*0.45*		

Table 3: Correlations between the model's production probability vs adult and children's data. Significant correlations highlighted in bold. CMS and KC didn't report significance level.

Correlation	Regular		Irregular	
r	Rule-based	Analogy	Rule-based	Analogy
246	**0.48**	**0.58**	**0.34**	0.00
500	**0.35**	**0.35**	0.25	0.02
1500	0.25	**0.27**	**0.34**	0.10
3000	**0.33**	**0.32**	**0.33**	0.09

Table 4: Correlations between the model's production probability vs the rule-based model and the analogy model. Significant correlations highlighted in bold.

model with 1500 verbs. The correlations with rule-based score is not significantly different from the analogy score for regulars. For irregulars, none of the models is correlated with the analogy model; models with 246, 1500 and 3000 verbs are significantly correlated with rule-based score. It seems that analogy score better correlates with regulars and rule-based score better correlates with irregulars. *Among models*: For regulars, the model with 246 verbs has the highest correlation with analogy score and rule-based score, and is significantly higher than model with 1500 verbs and 300 verbs. For irregulars, the correlations of rule-based score are not significantly different among models. **Summary:** Most of the models correlate with the rule-based model for both regulars and irregulars. This result shows that the neural network models are not completely incompatible with the rule-based theory. However, the models only correlate with the analogy model for regulars, but not for irregulars. This interesting dichotomy might suggest that the neural models may distinguish regulars and irregulars in processing. In addition, the model with 246 verbs also only correlate with children's data for irregular but not regular. This result might suggest that the mechanism to process irregulars for the model and children might be more closer to what rule-based model describes, therefore resulting in significant correlation.

4.1.4 Nonce verb output

We also look at the models' average production of the regular and the suggested irregular forms, as shown in Figure 1. Human speakers produce a variety of regulars and irregulars, as well as other forms not included in AH's report. However, for models with 500, 1500 and 3000 verbs, the models predominately produce the regular form for most of the verbs except for one verb: 'fleep'-/flɛpt/. Only the model with 246 verbs exhibit some variety of regular and irregulars in the prediction.

In addition, many 'other' forms the models produced are not human-like. Some common types include vowel change + ed, e.g. 'bize'/baɪz/ - /baʊzd/ or /bɔɪzd/, and consonant change, e.g. 'flidge' /flɪʤ/ - /flfʤ/.

4.2 Experiment 2: Evaluating on real verbs

In this experiment, we aim to conduct an error analysis on the models' real verb output to see if there's

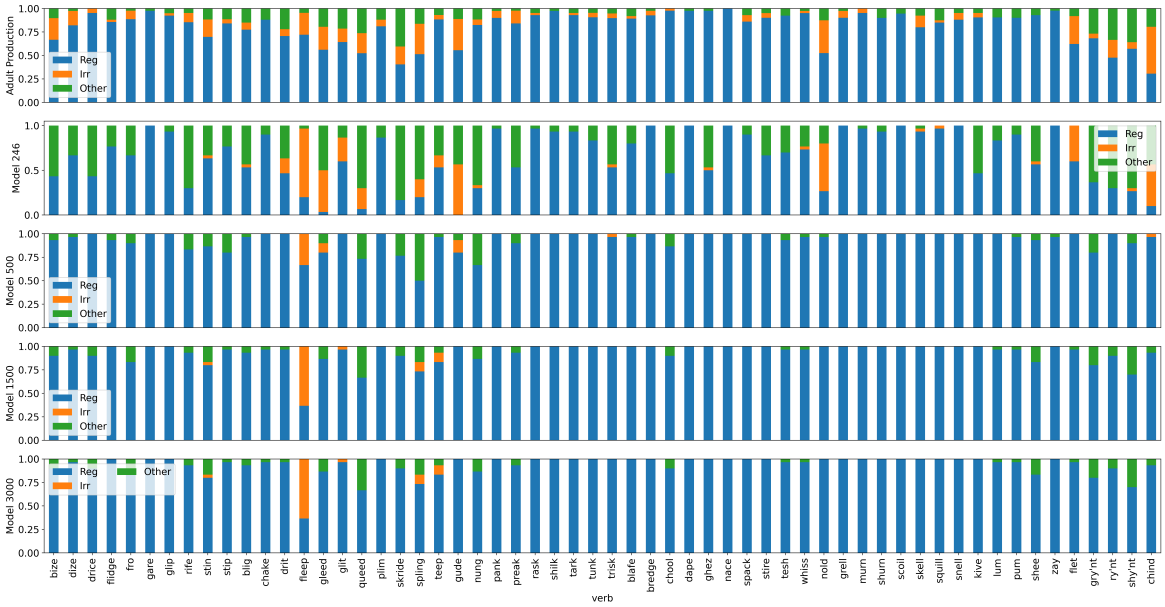

Figure 1: Percentage of regular, irregular and other predictions by humans and models

differentiation between regulars and irregulars and if the models make any overregularization errors.

First, we report the test accuracy on the real verb set, listed in Table 5. The large train size models (with 500, 1500 and 3000 verbs) reached near-perfect accuracy for the regular verbs and the small-size model's accuracy is poor. Also, all model's achieved some accuracy on irregular verbs.

	Regulars %			Irr
Size	/-d/	/-t/	/-d/	irr %
246	80 (5.4)	89 (4.2)	49 (8.8)	17 (4.6)
500	98 (1.7)	97 (1.7)	96 (3.3)	5 (3.2)
1500	99 (1.2)	98 (1.4)	99 (1.2)	13 (4.7)
3000	99 (1.2)	99 (1.3)	99 (2.2)	27 (3.6)

Table 5: Mean accuracy of test set with 170 verbs (standard deviations in brackets) averaged over 30 samples for each data size. There might be some overlapping in the training data and test data, since training data are generated randomly.

4.2.1 Distinction between regulars and irregulars

We analyzed all the errors made by each model with different data size and roughly divided them into 5 categories. **1. No change:** the model output is the same as the root, e.g. 'oversee': /oʊvərsi/ - */oʊvərsi/, 'teach': /tiʧ/ - */tiʧ/ **2. Plural /d/**: the model erroneously produced multiple /d/s at the end of the verb, e.g. 'withdraw': /wɪθdrɔ/ - */wɪθdrɔ*ddddddddd*/. **3. Allomorphy**: the model either output a wrong regular ending to a regular verb, e.g. 'bribe': /braɪb/ - */braɪb*t*/; or output a regular ending to an irregular verb, e.g. 'retell': /ritɛl/ - */ritɛl*d*/. **4. Consonant change**: the model erroneously changed the consonant in the root, e.g. 'secure': /sɪkj*ʊ*r/ - */sɪk*t*ʊrd/, 'force-feed': /fɔrs*f*i d/ - */fɔrs*t*id/'. **5. Vowel change**: the model erroneously changed the vowel in the root, e.g. 'rewrite': /rir*aɪ*t/ - */rir*oɪ*t/', 'giggle': /gɪgəl/ - */g*a*gəld/.

We tabulated each model's different types of error in contingency Table 6 and conducted chi-square analysis to test if there is association between error types and regularity. Since some cell numbers are lower than 5, we used Fisher's exact test instead of chi-square test. The p-value is significant for model with 246 verbs, 500 verbs and 1500 verbs, suggesting that these models make different errors for regulars and irregulars. There is no significant distinction in error types for regulars and irregulars in model with 3000 verbs, probably due to the low number of errors. The error type associations with regularity are different for model with 246, 500 and 1500 verbs, as shown in Table 7. All three models tend to make Plural /d/ and Allomor-

phy errors on irregulars. Model with 246 and 500 verbs tend to make No change and Vowel change errors on regulars. Model with 500 and 1500 verbs tend to make Consonant change errors on irregulars. The differences in the regular-irregular association might be explained the low number of errors on regulars in model with 500 and 1500 verbs.

Size	246		500		1500		3000	
	R	I	R	I	R	I	R	I
1	591	44	60	42	6	57	7	43
2	4	83	3	275	1	78	0	19
3	31	62	7	88	2	107	4	32
4	134	48	11	85	8	116	7	48
5	466	115	60	37	31	52	14	48
p	**<.001**		**<.001**		**<.001**		0.14	

p=Fisher's test p value, R=regular, I=irregular,
1=No change, 2=Plural /d/, 3=Allomorphy
4=Consonant Change, 5=Vowel Change

Table 6: Contingency table of the frequency of errors of different type in models with different size. The Fisher's exact p-value is significant for three models, highlighted in bold.

Size	246	500	1500
1.No change	Reg	Reg	Irr
2.Plural /d/	Irr	Irr	Irr
3.Allomorphy	Irr	Irr	Irr
4.Consonant Change	Reg	Irr	Irr
5.Vowel Change	Reg	Reg	Irr

Table 7: The different types of errors each model tend to make on regulars or irregulars

The distinction between regular error type and irregular error type is very interesting. We wonder how the model learned this distinction: is it learned based on the verb stem or the past tense forms? To further investigate this distinction, we trained 6 more models with only regular verbs with training size ranging from 500 - 3000 and tested it on the same real verb test set. Since there is no irregular verbs in the training data, we expect model to produce the regular past tense ('+ed') for the irregulars. The 6 models all have 100 accuracy on regulars and 0 accuracy in irregulars. However, we only found 2 '+ed' errors on the irregulars: 'deal': /dild/, 'retell':/ritɛld/. All the models produced Plural /d/ errors on the rest of the 18 irregular verbs. This result further confirms that the model learned the regular-irregular distinction, and suggests that the distinction is learned from verb stem.

4.2.2 Overregularization Errors on irregulars

We found all three types of overregularization errors in our model output, as listed in Table 8. In addition, the model also made many novel errors, such as incomplete suffix (e.g. rewrite - */riraɪtɪ/), double suffix (e.g. awake - */əweɪktd/) and truncation (e.g. stand - */stæn/). A more careful qualitative analysis on these errors should help us to understand more of the model's behavior.

Type	Examples
Stem+ed	deal - */dild/*, stick - */stɪkt/*
Past+ed	sink - */sæŋkt/*, awake - */əwəʊkt/*
Incorrect vowel change	swing-*/swæŋ/*, oversee-*/oversɛ/*

Table 8: Examples of overregularization errors made by models

5 Discussion

In this work, we showed that the transformer model is currently insufficient as a cognitive model, but exhibits some human-like characters. We found that all neural models have significant correlations with adult behavior's in both regulars and irregulars. The model with 246 verbs of the same distribution as parent's input correlates with children's irregular behavior, but not the regulars. The models correlate with rule-based model on regulars and with analogy model on irregulars. The dichotomy in correlations with cognitive models and children's data suggested that the model's behavior and children's behavior on irregular verbs are more closer to what rule-based theory describes. For nonce verb production, the model with 246 verbs show some variety as in human speakers, but such variety is not found in other models.

We also found overregularization errors the models make that are similar to children's errors. Although the models make many non-human like errors, we show that these errors exhibit a clear distinction between regulars and irregulars. The model possibly learned the regular-irregular distinction from the verb stem instead of the past tense forms. The error data also confirms that models mimic human behavior.

One important difference of our neural models and KC, CMS is that we manipulated the training data. We showed that model with small-size training data with high proportion of irregulars correlates better with human behavior and cognitive

models' score. However, the small-size model that replicates parents' verb distribution generally have lower accuracy than human children. If we can improve the accuracy without flooding the model with more training data, we could better demonstrate that neural networks can be good cognitive models.

To further evaluate neural networks, there are many other potential aspects that can be explored, such as a more careful error analysis, inflections in other languages, or visualizing hidden layers to help us understand what the neural networks learned. We hope that our evaluation could motivate more future explorations of neural networks as cognitive models.

References

Adam Albright and Bruce Hayes. 2003. Rules vs. analogy in english past tenses: A computational/experimental study. *Cognition*, 90(2):119–161.

R Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. The celex lexical database (release 2). *Distributed by the Linguistic Data Consortium, University of Pennsylvania*.

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

Lois Bloom. 1973. *One word at a time: The use of single word utterances before syntax*, volume 154. Walter de Gruyter.

Lois Bloom, Lois Hood, and Patsy Lightbown. 1974. Imitation in language development: If, when, and why. *Cognitive psychology*, 6(3):380–420.

Ryan P Blything, Ben Ambridge, and Elena VM Lieven. 2018. Children's acquisition of the english past-tense: Evidence for a single-route account from novel verb production data. *Cognitive Science*, 42:621–639.

Michael B Broe. 1993. *Specification theory: the treatment of redundancy in generative phonology*. Ph.D. thesis, University of Edinburgh.

Joan L Bybee and Carol Lynn Moder. 1983. Morphological classes as natural categories. *Language*, pages 251–270.

Maria Corkery, Yevgen Matusevych, and Sharon Goldwater. 2019. Are we there yet? encoder-decoder neural networks as cognitive models of english past tense inflection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3868–3877.

Roy Patrick Higginson. 1985. *Fixing: Assimilation in language acquisition*. Ph.D. thesis, Washington State University.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Elena Lieven, Dorothé Salomo, and Michael Tomasello. 2009. Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20(3):481–507.

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Michael Maratsos. 2000. More overregularizations after all: new data and discussion on marcus, pinker, ullman, hollander, rosen & xu. *Journal of Child Language*, 27(1):183–212.

Gary F Marcus, Steven Pinker, Michael Ullman, Michelle Hollander, T John Rosen, Fei Xu, and Harald Clahsen. 1992. Overregularization in language acquisition. *Monographs of the society for research in child development*, pages i–178.

Robert JC Maslen, Anna L Theakston, Elena VM Lieven, and Michael Tomasello. 2004. A dense corpus study of past tense and plural overregularization in english.

Steven Pinker and Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1-2):73–193.

Kim Plunkett and Virginia Marchman. 1991. U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition*, 38(1):43–102.

David. E. Rumelhart and James L. McClelland. 1986. *On Learning the Past Tenses of English Verbs*, page 216–271. Cambridge, MA, USA.

Jacqueline Sachs. 1983. Talking about the there and then: The emergence of displaced reference in parent-child discourse. *Children's language*, 4:1–28.

Wayne A Wickelgren. 1969. Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, 76(1):1.

Fei Xu and Steven Pinker. 1995. Weird past tense forms. *Journal of child language*, 22(3):531–556.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

A Appendix

Tokens		Parent's Regular			Parent's Irregular
Child	Files	*/-d/*	*/-t/*	*/-ɪd/*	*irr*
Adam[1]	18	18	18	3	36
Eve[1]	5	5	7	3	18
Sarah[1]	33	13	17	0	33
Peter[2]	14	1	3	0	8
Naomi[3]	20	9	9	4	27
Allison[4]	6	8	4	1	18
April[5]	2	5	5	1	17
Fraser[6]	90	83	44	17	62

1.Bloom (1973), 2.Bloom et al. (1974), 3.Sachs (1983),
4.Bloom (1973), 5. Higginson (1985),
6.Lieven et al. (2009)

Table 9: Summary of each parent's regular verb and irregular verb tokens