Cloze Distillation: Improving Neural Language Models with Human Next-Word Predictions

Tiwalayo N. Eisape[1] **Noga Zaslavsky**[1,2] **Roger P. Levy**[1]
[1]Department of Brain and Cognitive Sciences, [2]Center for Brains Minds and Machines
Massachusetts Institute of Technology
{eisape, nogazs, rplevy}@mit.edu

Abstract

Contemporary autoregressive language models (LMs) trained purely on corpus data have been shown to capture numerous features of human incremental processing. However, past work has also suggested dissociations between corpus probabilities and human next-word predictions. Here we evaluate several state-of-the-art language models for their match to human next-word predictions and to reading time behavior from eye movements. We then propose a novel method for distilling the linguistic information implicit in human linguistic predictions into pre-trained LMs: Cloze Distillation. We apply this method to a baseline neural LM and show potential improvement in reading time prediction and generalization to held-out human cloze data.

1 Introduction

Modern language models (LMs) demonstrate outstanding general-purpose command over language. The majority of these models acquire language by maximizing the in-context probability of each word in their training corpus (Figure 1), typically with a self-supervised objective. This simple corpus probability matching has resulted in models that learn impressive powers of both psychometric prediction (Frank and Bod, 2011; Fossum and Levy, 2012; Frank et al., 2015; Goodkind and Bicknell, 2018; Hale et al., 2018; van Schijndel and Linzen, 2018; Warstadt and Bowman, 2020; Wilcox et al., 2020) and language more generally (Devlin et al., 2019; Radford et al., 2019).

In humans, prediction may underlie both learning (Kuhl, 2004; Huang and Snedeker, 2013) and processing (Ryskin et al., 2020; Levy, 2008; Clark, 2013). Human linguistic prediction can be understood as not only lexical but also as taking place both above and below the word level (Federmeier and Kutas, 1999; Federmeier et al., 2002); parallel, i.e., predictive commitments are maintained over several linguistic units at once (Levy, 2008); and graded, i.e., commitment is licensed to varying degrees based on features of the linguistic unit being predicted. Rather than placing bets (Jackendoff, 1987) on which single word will come next, humans make many diffuse bets at multiple linguistic levels (e.g., syntactic, orthographic, lexical, etc.).

Surprisal theory (Hale, 2001; Levy, 2008) describes the utility of the approach taken by the human language processor, as lexical prediction is often an ill-constrained classification problem — for agents with very large vocabularies (LMs, humans), context is often not sufficiently constraining for high accuracy multiple, thousand-way classification decisions, but is typically constraining enough to accurately infer next-word features (such as part of speech, and semantic category). A large body of evidence demonstrates that these graded next-word predictions are reflected in human processing times (Ehrlich and Rayner, 1981; Demberg and Keller, 2008; Smith and Levy, 2013; Luke and Christianson, 2016) as well as neural responses (Kutas and Hillyard, 1980; Frank et al., 2015).

Corpus data are (imperfect) samples from the linguistic environment of a native speaker, and psycholinguistic data indicate that accurate prediction is important to efficient language comprehension. Under the principle of rational analysis (Anderson, 1990), it is thus to be expected that artificial language models trained on corpus data would correlate with human linguistic predictions and thus have good psychometric predictive accuracy. Nevertheless, past work (Smith and Levy, 2011) has suggested dissociations between corpus probabilities and human next-word estimates. Here, we further investigate this relationship using artificial language models and the most extensive corpus of sequential cloze completions that we are aware of: the Provo Corpus (Provo henceforth; Luke and

Proceedings of the 24th Conference on Computational Natural Language Learning, pages 609–619

https://doi.org/10.18653/v1/P17

Christianson, 2018).

First, we use Provo to test the psychometric performance of three state-of-the-art Transformer-based (Vaswani et al., 2017) LMs — XLNet (Yang et al., 2019), Transformer-XL (Dai et al., 2019), and GPT-2 (Radford et al., 2019) — alongside a smaller 2-layer LSTM (Hochreiter and Schmidhuber, 1997) trained on wikitext-103 (Merity et al., 2016), and a 5-gram LM baseline (Stolcke, 2002). We find that, while the Transformer models achieve the lowest perplexity on Provo and the best fit to the cloze data, the LSTM model provides the best account of reading times in terms of raw correlation. These findings show a dissociation between recapitulating corpus statistics and mimicking human language processing, operationalized here with reading times. That is, models that minimize perplexity on next-word prediction do not necessarily provide the best account of reading times. Second, based on these findings, we propose Cloze Distillation: a novel method for distilling linguistic information implicit in human cloze completions into pre-trained LMs. We apply this method to the LSTM model and show substantial improvement in reading time prediction and word frequency estimation, in addition to generalization to held-out human cloze data.

2 Human Cloze Predictions

The objective for most modern LMs is to compute a probability distribution over the model's vocabulary V for the likely next-word $x \in V$ at position i given the context $\mathbf{x}_{<i}$ consisting of the sequence of preceding words in the document. Similarly, as humans process language, they make constant and implicit linguistic predictions.

One commonly used measure of these predictions in humans is the Cloze task. In its original form (Taylor, 1953), the task involved masking a word or words in a source text passage and asking participants to provide words for the masked elements that would make the passage "whole again", a task structure adopted by contemporary masked language models (Devlin et al., 2019). In experimental psycholinguistics, however, the most common version of the Cloze task has involved presenting the beginning, or *prefix*, of a passage and having participants either complete it or provide the word that they think comes next (Figure 1), a task more closely matching that of autoregressive language models (Radford et al., 2019). In this paper, we focus on this latter type of Cloze task, which elicits samples from comprehenders' subjective next-word probability distributions (DeLong et al., 2005; Staub et al., 2015). For any given prefix, we can estimate the cloze distribution of a typical native speaker from pooled cloze responses across a large number of participants (Luke and Christianson, 2018), similar to how the fundamental output of an autoregressive language model is a vector of next-word probabilities.

2.1 The Provo Corpus

We use the Provo Corpus (Luke and Christianson, 2018) as our source of paired cloze completion and reading time data. The Provo Corpus derives from 55 paragraphs of text taken from sources including online news articles, popular science, and fiction. For each paragraph p, next-word cloze completions were elicited for each prefix $\mathbf{x}_{<i}$ for $i = 2, \ldots |p|$ (2,689 sentence prefixes total). Prefixes were presented to participants ($N = 470$) as a continuous multi-line text (Figure 1). This resulted in an average of 40 cloze responses with 15 unique continuations per prefix.

Additionally, Luke and Christianson (2018) collected eye movement data from eighty-four native speakers of American English as they read these 55 text passages, using a high-resolution SR Research EyeLink 1000 eye tracker.

The Provo cloze data, eye movement data, and the relationship between them are analyzed in detail in (Luke and Christianson, 2016). Luke and Christianson (2016) point out that while context is rarely constraining enough to facilitate exact next-word prediction, modal cloze responses often constitute *partial* matches to the target words. For example, given the prefix *With schools still closed, cars still buried and streets still ...*, the true continuation, *blocked*, has a cloze probability of only 0.07. But the overwhelming majority of cloze responses are partial fits to the correct word: 79% of the responses are verbs, and 72% are inflectional matches (ended with -ed), with the two most frequent responses being *closed* and *covered* (example from Luke and Christianson, 2018). In addition, they showed that cloze probabilities are highly predictive of reading times, adding to prior work showing a word's reading time is a function of its predictability in context (e.g., Smith and Levy, 2013).

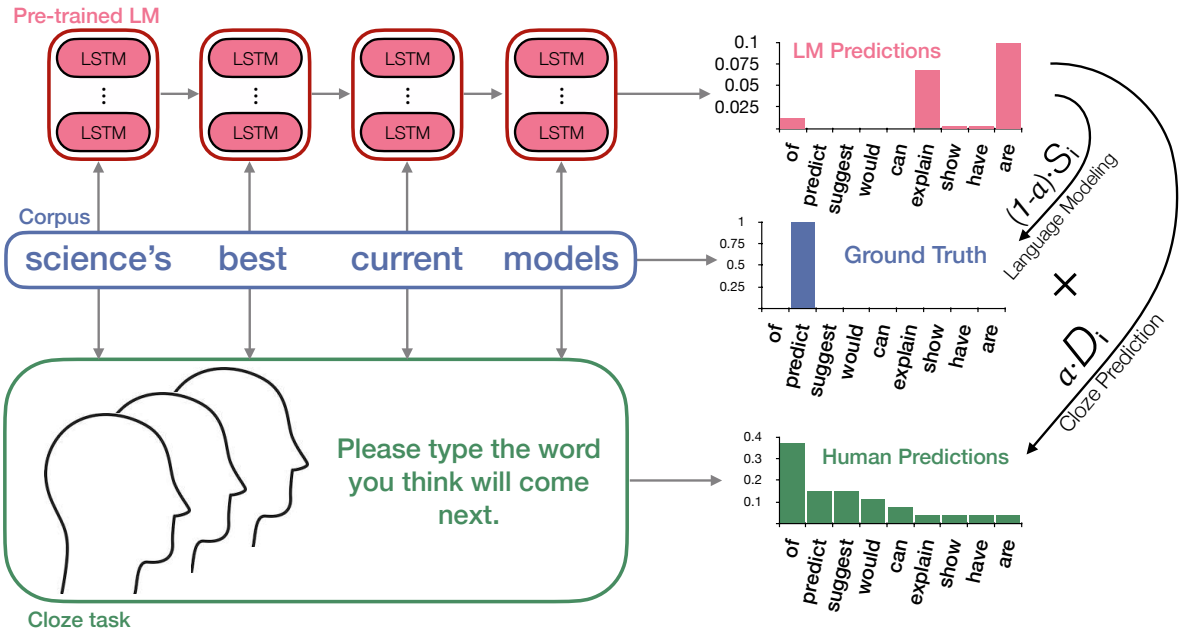

Figure 1: Illustration of the Cloze task and the Cloze Distillation objective. Given one of Provo's prefixes — in this example, one that ends in . . . *science's best current models*, where the true next word (ground truth) is *predict* — human subjects were prompted, as shown in the Cloze task box, to predict the word they thought was likely to follow. The Cloze Distillation loss is constructed by combining (1) the KL divergence D_i between the human cloze distribution and the LM's next-word distribution, and (2) the LM's predicted surprisal S_i of the true next word given the prefix.

3 Testing Language Models on Provo

The findings of Luke and Christianson (2016) highlight cloze as a useful test-bed for LMs. Specifically, a LM that employs predictions similar to those that underlie human language processing is expected to be a good model of human cloze responses. Therefore, we evaluate here a suite of LMs on their ability to match human cloze distributions. Additionally, we use the LMs' ability to predict reading times as a second measure of fit to human expectations, extending past work using LMs to predict reading times (Frank and Bod, 2011; Wilcox et al., 2020).

3.1 Models

We consider in our analysis the following LMs:

1. **5-gram**: N-gram model using a window size of 5 with Kneser-Ney smoothing, obtained via the SRILM language modeling toolkit (Stolcke, 2002).

2. **LSTM**: A standard 2-layer LSTM RNN implemented in PyTorch (Paszke et al., 2017), used here with 256 hidden units and word embedding size of 256, and trained on the wikitext-103 corpus (Merity et al., 2016) via a next-word prediction task (40 epochs, batch size = 40, learning rate = 20).

3. **GPT-2**: A Transformer-based LM trained on the WebText corpus (Radford et al., 2019).

4. **Transformer-XL** (TXL; Dai et al., 2019): A Transformer-based LM with a segment level recurrence mechanism and relative positional embeddings trained on wikitext-103.

5. **XLNet** (Yang et al., 2019): A Transformer-based LM trained with a permutation language modeling objective as well as a segment level recurrence mechanism and relative positional embeddings. Training data consists of ∼30 billion tokens across 6 different copora.

We use the `LMzoo` python package (Gauthier et al., 2020) to access the 5-gram model, and the HuggingFace `transformers` python package (Wolf et al., 2019) for accessing Transformer models (`gpt2-large`, `transfo-xl-wt103`, and `xlnet-large-cased` respectively). These Transformer models use subword tokens (Sennrich et al., 2016); we defined word probabilities for these models as the joint probability of the subword tokens comprising the word given the context.

Model	$\langle D_i \rangle$	$\langle \tau_i \rangle$	$\langle S_i \rangle$	F_{intr}	F_{base}	ρ_{gaze}	ρ_{freq}
Cloze	NA	NA	3.99 ± 2.60	198.10	30.90	0.36	-0.43
GPT-2	2.30 ± 1.57	-0.57 ± 0.004	6.11 ± 5.00	252.70	46.11	0.40	-0.46
XLNet	2.39 ± 1.68	-0.58 ± 0.005	6.39 ± 5.70	260.50	46.08	0.41	-0.48
TXL	3.27 ± 1.92	-0.47 ± 0.005	8.09 ± 5.50	238.30	30.54	0.39	-0.50
LSTM	3.74 ± 1.86	-0.39 ± 0.006	8.58 ± 4.90	361.20	41.47	0.47	-0.63
5-gram	3.89 ± 1.84	-0.20 ± 0.007	12.48 ± 7.00	161.00	16.72	0.31	-0.41

Table 1: Evaluation of LMs on Provo reveals a dissociation between performance on next-word prediction and psychometric measures that reflect human language processing. F_{intr} and F_{base} show the F-test statistics (Section 3.2.2) against various baseline predictors. ρ_{gaze} and ρ_{freq} show correlation with gaze and frequency respectively (Pearson's ρ). $\langle D_i \rangle$ is average KL-divergence between the empirical cloze distribution and the LM's distributions; $\langle \tau_i \rangle$ is rank correlation between down-sampled model surprisals and surprisal values based on the empirical cloze probabilities; $\langle S_i \rangle$ is average surprisal over the text in Provo; all standard deviations are computed by paragraph.

3.2 Metrics

We use several metrics to evaluate the fit of our models to human reading times and cloze responses. We discuss and motivate them in the following section.

3.2.1 Cloze Responses

We use two measures to evaluate the performance of each model on human cloze data. First, we measure the deviation between the empirically estimated cloze distribution, $P_{\text{cloze}}(x|\mathbf{x}_{<i})$, where x is a potential next-word at position i in a document[1] and the model's next-word distribution, $P_{\text{model}}(x|\mathbf{x}_{<i})$, using the Kullback-Leibler (KL) divergence:

$$D_i \equiv D\left[P_{\text{cloze}}(x|\mathbf{x}_{<i}) \| P_{\text{model}}(x|\mathbf{x}_{<i})\right] \quad (1)$$
$$= \sum_{x \in V} P_{\text{cloze}}(x|\mathbf{x}_{<i}) \log \frac{P_{\text{cloze}}(x|\mathbf{x}_{<i})}{P_{\text{model}}(x|\mathbf{x}_{<i})} \, .$$

While the KL divergence is a natural measure for comparing distributions, it is potentially limited for our purposes due to the sparsity of the cloze data. To address this, we also consider Kendall's Tau correlation coefficient, which may be more robust to estimation errors resulting from small sample effects. Specifically, we consider Kendall's Tau correlation between LM surprisals and surprisals estimated form human cloze data, denoted here by $\tau_i \equiv \tau\left[P_{\text{cloze}}(x|\mathbf{x}_{<i}), P_{\text{model}}(x|\mathbf{x}_{<i})\right]$.

To further evaluate the models' ability to mimic cloze responses and to control for the sparsity of the human cloze data, we simulated a cloze task experiment with our LMs. For each LM, we generated 40 cloze responses[2] per prefix $\mathbf{x}_{<i}$ in Provo by sampling from $P_{\text{model}}(x|\mathbf{x}_{<i})$. We repeated this experiment 50 times for each model. The results were similar in both the down-sampling and without-down-sampling conditions, and we report only the down-sampling condition in Table 1.

3.2.2 Reading Times

We use *gaze duration* during first-pass reading as our measure of reading times, which is the amount of time a reader's eyes spend on a word the first time they fixate it (Rayner, 1998; if a reader fixates a word to the right before fixating the word in question, the word has been "skipped" and there is no valid gaze duration). It is well established that gaze duration captures a wide variety of cognitive processes during real-time language-comprehension, including the relationship between a word and the context in which it appears (Staub, 2011).

We evaluate the ability of a LM to account for human reading times based on their predicted surprisal values,

$$S_i \equiv -\log_2 P_{\text{model}}(x_i|\mathbf{x}_{<i}) \, , \quad (2)$$

as it has been previously shown to capture several characteristics of human language comprehension and pattern with reading times (Smith and Levy, 2013; Wilcox et al., 2020). Similarly, we define cloze surprisals by taking the negative log of the empirical cloze probabilities[3], i.e.,

[1] As participants in Luke and Christianson (2018) were given only within-paragraph context when prompted for each cloze response, each paragraph constitutes a unique document in our analysis.

[2] We generated 40 responses because most prefixes in Provo had at least 40 responses provided by participants.

[3] We use the cloze probability estimates from Luke and Christianson (2018)'s 'Orthographic Match Model' – a logit mixed-effects model including only random by-word intercepts. These estimates are nearly perfectly correlated with the relative frequency estimate of cloze ($\rho = .999$), but crucially

$-\log_2 P_{\text{cloze}}(x_i|\mathbf{x}_{<i})$. We then measure Pearson's correlation ρ between reading times and surprisal values. In addition, we use ANOVA tests to measure the models' predictive capacities beyond standard baseline predictors of reading time (Howes and Solomon, 1951; Kliegl et al., 2006; Leyland et al., 2013) — log word frequency and word length. That is, for each model (either an LM or the cloze distribution), we enter its surprisal values into a linear mixed-effects model (LME) along with the baseline predictors, and measure their contribution by computing the F-test statistic between the full LME and an LME where model surprisals are ablated out. In the case of $\boldsymbol{F}_{\textbf{base}}$ the baseline predictors were frequency, length, and their interaction. In the case of $\boldsymbol{F}_{\textbf{intr}}$ the baseline predictors were simply random by-word intercepts. We use both word frequencies estimated from the Corpus of Contemporary American English (COCA; Davies, 2010) and from wikitext-103 (Merity et al., 2016) in our analysis. As the results of our analyses were qualitatively the same in both conditions we report only results from COCA in the analyses to follow.

3.3 Results

The main results of evaluating the LMs on Provo are summarized in Table 1. First, averaging the KL divergence and suprisals values over word positions i in Provo (that is, $\langle D_i \rangle$ and $\langle S_i \rangle$ respectively), shows that the ability of LMs to predict human cloze responses tracks with their language modeling performance. This pattern is also reflected in Kendall's τ correlation between model surprisals and surprisals constructed from the human cloze distribution. At the same time, Table 1 reveals a dissociation between next-word prediction, reflected by $\langle S_i \rangle$, and human language processing, as reflected in reading times. Specifically, the LSTM model, which does not perform as well as the Transformer-based LMs in next-word prediction on Provo, as reflected in its higher $\langle S_i \rangle$, exhibits superior ability in predicting reading times, as measured in ρ_{gaze} and F_{intr}. This result is similar to that of Merkx and Frank (2020), who found that Gated Recurrent Unit networks outperformed Transformer models with lower perplexity in predicting gaze duration.

We note that when predicting reading times not only from the model's surprisal values, but also using the baseline predictors (word frequency and length), the LSTM model no longer outperforms the Transformer-based models (Table 1, F_{base}). Nonetheless, it is striking that the LSTM model, which is much smaller than the Transformer-based models and was trained on much less data, achieves the best performance in predicting reading times without the baseline predictors.

3.4 Intermediate Conclusions

Past work shows that human predictions systematically diverge from corpus probabilities (Smith and Levy, 2011). Our analysis extends these findings by testing current state-of-the-art LMs trained on much larger datasets, and showing that, while better estimates of corpus probabilities may yield better models of human next-word predictions, there does not seem to be a strict positive correlation between the ability to approximate corpus probabilities and the ability to predict human reading times, as evidenced by models with higher $\langle S_i \rangle$ being on-par and even better at predicting reading times compared to models with lower $\langle S_i \rangle$.

Recent studies (Ettinger, 2020; Hao et al., 2020; Jacobs and McCarthy, 2020) have found similar trends when comparing LMs to cloze data. Hu et al. (2020) also found only a loose relationship between perplexity (a monotonic function of $\langle S_i \rangle$) and syntactic generalization, adding to a growing body of evidence suggesting that while optimizing for corpus probabilities can create somewhat psycholinguistically-enabled language models (Linzen et al., 2016; Futrell et al., 2019; Hu et al., 2020), there may be a dissociation between corpus probabilities and human expectations.

4 Cloze Distillation

Here, we show how to leverage these findings to improve the ability of LMs to match human expectations, providing more appealing neural language models for human language processing. To this end, we propose Cloze Distillation: a method for using human next-word predictions as learning targets together with corpus statistics within a knowledge distillation framework.

4.1 Knowledge Distillation

Knowledge distillation (Buciluundefined et al., 2006; Ba and Caruana, 2014; Hinton et al., 2015) is a technique of imbuing knowledge from a teacher model into a student model by training the student to make the same predictions as the teacher. Typ-

do not include cloze probabilities of zero (which would yield infinite surprisal).

ically deployed as a form of model compression, knowledge distillation is useful for those looking to deploy insights from one or more complicated models into a single smaller model. Recently, knowledge distillation has also proven useful to cognitive scientists in creating low-dimensional neural network cognitive models (Schaeffer et al., 2020). When humans are used as the 'teacher' this can be seen as a specific case of a more general cognitive modeling strategy, task-based modeling.

4.2 The Cloze Distillation Objective

Knowledge distillation has proven its usefulness in NLP where researchers have distilled knowledge from very large and/or syntactically aware language models into naive models showing it is possible to transfer even subtle linguistic preferences from teacher to student (Kim and Rush, 2016; Kuncoro et al., 2019; Sanh et al., 2020; Kuncoro et al., 2020).

We take inspiration from this work and leverage the general framework both as a method for distilling knowledge from a 'teacher' with desirable linguistic biases (humans in our case) and as a tool for cognitive modeling by using empirical cloze distributions P_{cloze} as target distributions in a knowledge distillation framework.

We follow this approach to arrive at the following loss function for Cloze Distillation (CD):

$$L_i = \alpha D_i - (1-\alpha) S_i \,. \tag{3}$$

That is, for each context $\mathbf{x}_{<i}$ we compute the CD loss by linearly interpolating D_i, the KL divergence between the distributions of the human teacher and the student model as defined in equation (1), with an autoregressive language modeling objective that places unit probability mass on the true next-word, formally defined by S_i in equation (2). Thus, CD fine-tunes LMs to predict the next word in the document while simultaneously producing a distribution over next-words that mirrors the empirical human cloze distribution for that context. This process is illustrated in Figure 1.

To evaluate the utility of the human cloze data, we vary the values of α from $\alpha = 0$, which corresponds to pure next-word prediction driven fine-tuning, to $\alpha = 1$, which corresponds to pure cloze-prediction based fine-tuning.

4.3 Cloze-Distilled LSTM

To begin to evaluate the CD paradigm, we apply it to the LSTM from Section 3 by fine-tuning this model using the CD objective over Provo. To test generalization and utilize the full corpus, we use a k-fold cross-validation scheme with $k = 55$, the number of paragraphs in Provo where humans are provided the full preceding paragraph as context. That is, each fold consists of data from one paragraph in the Provo dataset. We use 100 epochs for training. We provide our LM with the same context as humans, up to the beginning of the current paragraph.

Additionally, we vary α to test the utility of our cloze data and cross-validated separately for each value of α in the range $[0, 1]$, sampled at intervals of 0.05. This resulted in 1,155 unique models for testing. We wish to emphasize that even utilizing the entire Provo corpus via cross-validation, we are left with only 2685 training samples, which is minuscule with respect to the model's pre-training data (roughly 100 million samples). We refer to the resultant model as cloze-distilled LSTM (CD-LSTM).

4.4 Results

After fine-tuning on the CD objective, we note several interesting adaptions in model behavior. These mainly include significant improvement over the standard LSTM baseline in predicting human reading times and cloze distributions (Figure 2). We also discuss improvements in next-word prediction performance over Provo (Figure 3).

4.4.1 Reading times

Psychometric predictive capacity is starkly improved with Cloze Distillation, and the strength of the effect scales with α. This can be seen in Figure 2, which shows the statistical comparison of the CD-LSTM for varying levels of α. We add another model comparison designed to isolate the ability of CD-LSTM to predict reading times above the standard LSTM (Figure 2a). Specifically, we enter CD-LSTM's surprisals into an LME along with baseline predictors and surprisals from the standard LSTM and compute the F-test statistic against a LME with CD-LSTM surprisal ablated out.

CD-LSTM exhibits a significant improvement with α in its ability to predict reading times above the non-fine-tuned model (Figure 2a), as well as

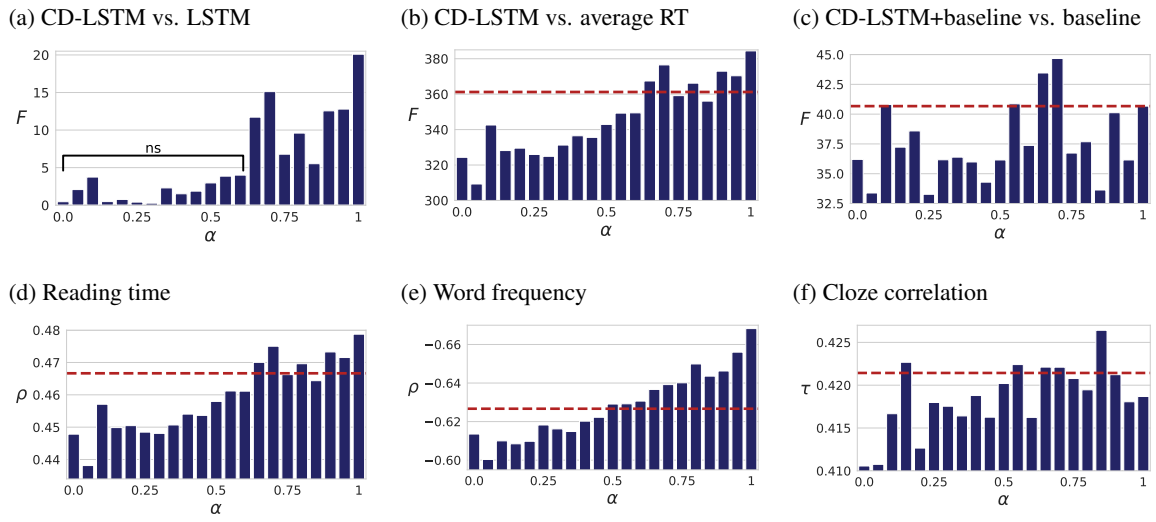

Figure 2: Results for CD-LSTM and the LSTM model (without any fine tuning) show that Cloze Distillation yields substantial improvement across several psychometric measures. Panels (a)-(c) show changes in F statistics as a function of α for three LME comparisons, and panels (d)-(f) show changes in three correlational measures. Dashed lines in panels (b)-(f) show the performance of the LSTM model. (a) LME based on CD-LSTM's surprisals outperforms the LME based on the LSTM's surprisals for most values of α (not significant for $\alpha < 0.65$). (b) LME based on CD-LSTM's surprisals outperforms the null (intercept only) model, and this performance generally improves with α. (c) LME based on CD-LSTM's surprisals with the baseline factors (word frequency and length) outperforms the baseline-only LME for several values of α. (d) Pearson's correlation between CD-LSTM's surprisals and reading times. (e) Pearson's correlation between CD-LSTM's surprisals and word frequencies. (f) Kendall's τ correlation between CD-LSTM's surprisals and human cloze surprisals.

improvements over an intercept-only model (Figure 2b) and baseline-only (Figure 2c). Correlation with reading time and CD-LSTM's surprisal also steadily increases with α (Figure 2d). These findings suggest that, as we postulate, Cloze Distillation is a useful paradigm for extracting the information about human linguistic expectations that is implicit in human cloze predictions and incorporating it into LMs.

4.4.2 Cloze

We report improvements in predicting held out cloze data, where $\langle D_i \rangle$ is decreased from 3.8 (at $\alpha = 0$) to 3.6 (at $\alpha = 0.65$) (Figure 3). τ correlation also exceeds that of the baseline model for several values of α (though there does not seem to be a consistent trend across α-s).

This result is intriguing as it implies that the requisite information for computing cloze distributions is learned over fine-tuning. Furthermore, we see a peak at $\alpha = 0.65$ and not at $\alpha = 1$, which suggests that in training LMs to predict cloze data, some signal from next-word prediction remains vital.

4.4.3 Language modeling

In addition to improved performance on our human language processing benchmarks, we see a robust increase in language modeling performance for most values of α, as evidenced by average surprisal over Provo (Figure 3). We note, the standard deviation in $\langle S_i \rangle$ for our LSTM over Provo was 1.86 bits (Table 1). The improvements we see are less than this deviation, and are thusly below the level of significance, though we do see a consistent trend in α. This effect is most substantial for intermediate values of α, suggesting that a combination of human knowledge and next-word prediction improves relative to either one of these factors on its own. This indicates that both parts of the loss function (ground truth next-words, human cloze) provide useful information for predicting text that is not entirely overlapping.

This is interesting given the low $\langle S_i \rangle$ of human cloze data. The fact that humans can contend with large language models trained explicitly on next-word prediction even on subsets of text, together with our Cloze Distillation results suggests there is linguistic information in human cloze that can be harnessed by LMs to subserve general language

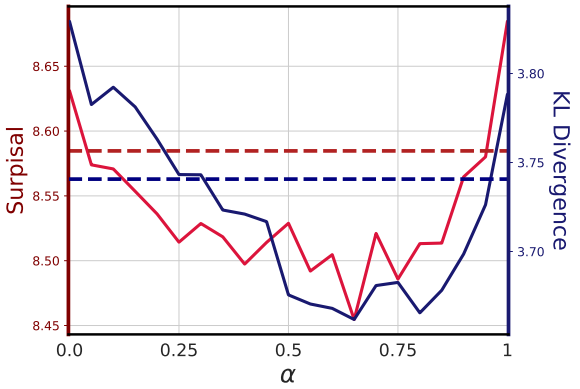

Figure 3: Average Surprisal (left) and KL divergence (right) over Provo as a function of the distillation interpolation coefficient α. Dashed lines show LSTM performance before fine-tuning.

modeling and is disjoint from the information accessible in corpus probability (Smith and Levy, 2011).

4.4.4 Frequency

We also note that as α increases, the CD-LSTM next-word predictions exhibit increased correlation with frequency (Figure 2e), suggesting that cloze distilled LMs may learn to better predict frequent words. This is interesting as a proof of concept that Cloze Distillation distills information implicit in cloze into language models as previous work (Smith and Levy, 2011) has shown human cloze is skewed toward more frequent words, relative to corpus probability.

5 Conclusion

Our analyses provide further evidence of a misalignment between language model estimates and human expectations. The method we provide: Cloze Distillation, demonstrates that shifting training incentives away from corpus probability toward psycholinguistic task-based modeling can result in better cognitive models and better language models. Still, given several of our models predict reading times beyond the cloze data collected in Provo (Table 1) there are several possible explanations for the effect Cloze Distillation has on language model performance.

One is that the Cloze task produces data that are a more faithful reflection of the expectations deployed in human reading and are thus able to guide the models toward a fundamentally more human-like set of expectations – despite being under-sampled. If this is true and human subjective next-word estimates also provide signal about next-word probabilities across corpora (reflecting the implicit knowledge speakers have learned about the statistics of their language), this would explain why Cloze Distillation improves next-word prediction accuracy on a new corpus (Provo).

Another possibility is that the models we survey are fundamentally better than the cloze data at capturing the human expectations deployed in reading. Though this would not explain the boost in performance we see in reading time prediction with Cloze Distillation, because several of our models predict reading times better than the cloze data itself, this can not yet be ruled out. We leave the further exploration of this to future work as larger-scale collection of human cloze data allows.

That said, the fact that we were able to induce appreciable adaptions in model behavior with such little data highlights the richly orienting information available in even noisy human predictions. Though it is unclear how language users learn to make such sophisticated predictions (we provided this information to our model with direct supervision), our model's ability to learn from such small scale data highlights the potential utility of such predictions in a language acquisition setting — it seems that human predictions are strong enough to significantly bolster the signal in raw linguistic input abetting extensive adaption from relatively little data.

As of now, the current dataset's scale restricts Cloze Distillation to use as a fine-tuning method. Furthermore, we use simple LSTMs to perform a detailed analysis of Cloze Distillation with dense sampling in α and thorough cross-validation. It is possible that deploying Cloze Distillation during pre-training in large models (e.g., Transformers) could result in models better able to learn the word features humans demonstrate knowledge of in their cloze responses and we leave the exploration of this to future work as well.

Methods such as Cloze Distillation provide an avenue forward for psycholinguists interested in taking LMs seriously as candidate models of human language processing and to natural language processing researchers interested in reverse engineering and deploying insights from human sentence processing. Cloze Distillation highlights these goals as potentially mutually-reinforcing.

6 Acknowledgements

TNE was supported by the GEM consortium and the MIT Dean of Sciences Fellowship. NZ was supported by an MIT BCS Fellowship in Computation. RPL was supported by NSF grant IIS1815529, a Google Faculty Research Award, and a Newton Brain Science Award. We thank Robert Chen for helping collect model surprisals, as well as Peng Qian and Jon Gauthier for helpful discussions.

References

John R. Anderson. 1990. *The Adaptive Character of Human Thought*. Hillsdale, NJ: Lawrence Erlbaum.

Jimmy Ba and Rich Caruana. 2014. Do Deep Nets Really Need to be Deep? In *Advances in Neural Information Processing Systems 27*, pages 2654–2662.

Cristian Buciluundefined, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model Compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD, page 535–541, New York, NY, USA. Association for Computing Machinery.

Andy Clark. 2013. Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. *Brain and Behavioral Sciences*, 36(3):181–204.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Mark Davies. 2010. The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English. *Literary and Linguistic Computing*, 25(4):447–464.

Katherine A DeLong, Thomas P Urbach, and Marta Kutas. 2005. Probabilistic Word Pre-Activation During Language Comprehension Inferred from Electrical Brain Activity. *Nature Neuroscience*, 8(8):1117–1121.

Vera Demberg and Frank Keller. 2008. Data from Eye-tracking Corpora as Evidence for Theories of Syntactic Processing Complexity. *Cognition*, 109(2):193–210.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Susan F. Ehrlich and Keith Rayner. 1981. Contextual Effects on Word Perception and Eye Movements During Reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641 – 655.

Allyson Ettinger. 2020. What BERT is not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Kara D Federmeier and Marta Kutas. 1999. A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language*, 41(4):469–495.

Kara D Federmeier, Devon B McLennan, Esmeralda De Ochoa, and Marta Kutas. 2002. The Impact of Semantic Memory Organization and Sentence Context Information on Spoken Language Processing by Younger and Older Adults: an ERP Study. *Psychophysiology*, 39(2):133–146.

Victoria Fossum and Roger Levy. 2012. Sequential vs. Hierarchical Syntactic Models of Human Incremental Sentence Processing. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 61–69, Montréal, Canada. Association for Computational Linguistics.

Stefan L Frank and Rens Bod. 2011. Insensitivity of the Human Sentence-processing System to Hierarchical Structure. *Psychological Science*, 22(6):829–834.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP Response to the Amount of Information Conveyed by Words in Sentences. *Brain and Language*, 140:1–11.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural Language Models as Psycholinguistic Subjects: Representations of Syntactic State. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An Online Platform for Targeted Evaluation of Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Adam Goodkind and Klinton Bicknell. 2018. Predictive Power of Word Surprisal for Reading Times is a Linear Function of Language Model Quality. In

Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018), pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. 2018. Finding Syntax in Human Encephalography with Beam Search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2727–2736, Melbourne, Australia. Association for Computational Linguistics.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic Predictions of People Perusing: Evaluating Metrics of Language Model Performance for Psycholinguistic Modeling. *arXiv preprint arXiv:2009.03954*.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Keural Network. In *Deep Learning and Representation Learning Workshop at NuerIPS*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

D.H. Howes and R.L. Solomon. 1951. Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology*, 41(6):401—410.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A Systematic Assessment of Syntactic Generalization in Neural Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Yi Ting Huang and Jesse Snedeker. 2013. The Use of Lexical and Referential Cues in Children's Online Interpretation of Adjectives. *Developmental Psychology*, 49(6):1090–1102.

Ray Jackendoff. 1987. *Consciousness and the Computational Mind*, volume 356. The MIT Press, Cambridge, MA.

Cassandra L. Jacobs and Arya D. McCarthy. 2020. The Human Unlikeness of Neural Language Models in Next-word Prediction. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, page 115, Seattle, USA. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Reinhold Kliegl, Antje Nuthmann, and Ralf Engbert. 2006. Tracking the Mind During Reading: The Influence of Past, Present, and Future Words on Fixation Durations. *Journal of Experimental Psychology*, 135:12–35.

Patricia K Kuhl. 2004. Early Language Acquisition: Cracking the Speech Code. *Nature Reviews Neuroscience*, 5(11):831–843.

Adhiguna Kuncoro, Chris Dyer, Laura Rimell, Stephen Clark, and Phil Blunsom. 2019. Scalable Syntax-Aware Language Models Using Knowledge Distillation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3472–3484, Florence, Italy. Association for Computational Linguistics.

Adhiguna Kuncoro, Lingpeng Kong, Daniel Fried, Dani Yogatama, Laura Rimell, Chris Dyer, and Phil Blunsom. 2020. Syntactic Structure Distillation Pretraining for Bidirectional Encoders. *arXiv preprint arXiv:2005.13482*.

Marta Kutas and Steven A. Hillyard. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Louise-Ann Leyland, Julie A. Kirkby, Barbara J. Juhasz, Alexander Pollatsek, and Simon P. Liversedge. 2013. The Influence of Word Shading and Word Length on Eye Movements During Reading. *Quarterly Journal of Experimental Psychology*, 66(3):471–486. PMID: 21988376.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Steven G. Luke and Kiel Christianson. 2016. Limits on Lexical Prediction During Reading. *Cognitive Psychology*, 88:22 – 60.

Steven G. Luke and Kiel Christianson. 2018. The Provo Corpus: A Large Eye-Tracking Corpus with Predictability Norms. *Behavior Research Methods*, 50(2):826–833.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer Sentinel Mixture Models. *arXiv preprint arXiv:1609.07843*.

Danny Merkx and Stefan L Frank. 2020. Comparing Transformers and RNNs on Predicting Human Sentence Processing Data. *arXiv preprint arXiv:2005.09471*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam

Lerer. 2017. Automatic Differentiation in PyTorch. *Neural Information Processing Systems Autodiff Workshop*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9.

Keith Rayner. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, 124(3):372–422.

Rachel Ryskin, Roger P Levy, and Evelina Fedorenko. 2020. Do Domain-General Executive Resources Play a Role in Linguistic Prediction? Re-evaluation of the Evidence and a Path Forward. *Neuropsychologia*, 136:107258.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBert, a Distilled Version of BERT:Smaller, Faster, Cheaper and Lighter. *arXiv preprint arXiv:1910.01108*.

Rylan Schaeffer, Mikail Khona, Leenoy Meshulam, and Ila Rani Fiete. 2020. Reverse-engineering Recurrent Neural Network Solutions to a Hierarchical Inference Task for Mice. *bioRxiv*.

Marten van Schijndel and Tal Linzen. 2018. Modeling Garden Path Effects without Explicit Hierarchical Syntax. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, pages 2600–2605, Austin, Texas. Cognitive Science.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Nathaniel Smith and Roger Levy. 2011. Cloze but no Cigar: The Complex Relationship between Cloze, Corpus, and Subjective Probabilities in Language Processing. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 33(33).

Nathaniel J Smith and Roger Levy. 2013. The Effect of Word Predictability on Reading Time is Logarithmic. *Cognition*, 128(3):302–319.

Adrian Staub. 2011. Word Recognition and Syntactic Attachment in Reading: Evidence for a Staged Architecture. *Journal of Experimental Psychology. General*, 140(3):407.

Adrian Staub, Margaret Grant, Lori Astheimer, and Andrew Cohen. 2015. The Influence of Cloze Probability and Item Constraint on Cloze Task Response Time. *Journal of Memory and Language*, 82:1 – 17.

Andreas Stolcke. 2002. SRILM - an Extensible Language Modeling Toolkit. In *Seventh international conference on spoken language processing*.

W. L. Taylor. 1953. "Cloze Procedure": A New tool for Measuring Readability. *Journalism Quarterly*, 30:415.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Alex Warstadt and Samuel R Bowman. 2020. Can Neural Networks Acquire a Structural Bias from Raw Linguistic Data? In *Proceedings of the 2020 Conference of the Cognitive Science Society*, pages 1737–1743.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. 2020. On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. In *Proceedings of the 2020 Conference of the Cognitive Science Society*, pages 1707–1713.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M Rush. 2019. HuggingFace's Transformers: State-of-the-Art Natural Language Processing. *arXiv e-prints*, page arXiv:1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in neural information processing systems*, pages 5753–5763.