Probing Reduplication Semantics with Neural Language Models

All code for this project is available on GitHub.

1 Introduction

1.1 Reduplication in Mandarin Chinese

Reduplication, most simply, is the doubling of any element to convey meaning. Often reduplication is defined in terms of phonological identity - whether identical in whole or in part - between the base and reduplicant (Marantz, 1982). Reduplication almost always carries a particular meaning (Inkelas and Zoll, 2005). It is a common construction cross-linguistically, as it is a simple and intuitive word formation strategy (Brdar, 2013).

Mandarin Chinese allows for reduplication of nouns, classifiers, adjectives, and verbs. There is a specific proposal in a series of papers by Melloni, Arcodia, and Basciano (henceforth MAB), which includes the form-meaning mapping for verbal reduplication in Mandarin (Arcodia et al., 2014, 2015; Basciano and Melloni, 2017; Melloni and Basciano, 2018).

They posit two different types of reduplication in Mandarin. Diminishing reduplication, which is characterized by inducing diminishing semantics, only applies to stems as opposed to roots, and takes the form ABAB for disyllabic verbs (Arcodia et al., 2014) and AA for monosyllabic verbs. Increasing reduplication, on the other hand, is associated with increasing semantics, applies to roots, and uses the AABB reduplication pattern. This analysis is centered on a clear form-meaning mapping between two types of inputs, two forms of reduplication, and two types of meanings, which is summarized in Table (1).

Input	Meaning type	Reduplication pattern	Domain
stem	diminishing semantics	ABAB, AA patterns	syntax
root	increasing semantics	AABB pattern	morphology

Table 1: Relationship described by Melloni and Basciano (2018)

1.2 Language Models as Psycholinguistic Subjects

Having language models complete psycholinguistic tasks and model human behavior is a relatively recent development. Already there are a variety of tasks which have been attempted, including grammaticality judgements (Linzen et al., 2016; Lau et al., 2017; Marvin and Linzen, 2018; Ek et al., 2019; Warstadt et al., 2019); structural priming (Prasad et al., 2019; Misra et al., 2020; Sinclair et al., 2022); language processing and prediction (Michaelov and Bergen, 2020); among others.

One motivation for this approach is that psycholinguistic stimuli are constructed by experts, and thus may avoid some of the issues that come from prompt engineering; however, due to the small sample size, it may be challenging to have sufficient statistical power to effectively proble a language model (Li et al., 2022).

2 Models

2.1 Model Architectures

2.1.1 Unidirectional Transformer Models

I conducted experiments on four types of unidirectional transformer models: Bloom (BigScience Workshop, 2022), XGLM (Facebook, 2022; Lin et al., 2021), Chinese GPT-2 (Chinese Knowledge and Information Processing, 2020), and GPT-3. These models differed mainly in the data that they were trained on. Bloom and XGLM were trained on multiple languages, including Mandarin Chinese. Chinese GPT-2 was trained exclusively on Mandarin Chinese data. GPT-3, on the other hand, was only trained on English data.

For GPT-3, I tested four engines: davinci, text-davinci-001, text-davinci-002, and text-davinci-003. I used the default query settings provided for the OpenAI API, which included temperature setting of 0.8 and a max length of 200.

For the Bloom, XLGM, and Chinese GPT-2 models, I used the pipeline function from HuggingFace to generate text. I used the AutoModelforCausalLM and AutoTokenizer functions to select the appropriate model and tokenizer for each model.

2.1.2 Bidirectional Transformer Models

I did not implement bidirectional transformer models for this experiment for two reasons. First, I did not find a Chinese BERT model available. Second, bidirectional transformer models do not have as good of prompt-based learning capabilities as unidirectional models (Patel et al., 2022). Patel et al. (2022) propose a method, Sequential Autoregressive Prompting, that not only allows bidirectional models to follow instructions, but furthermore to outperform the unidirectional models. This may be a promising direction for future experiments.

2.2 Training/Model Size

It has been demonstrated that model size may not always positively correlate with high performance for all task. Li et al. (2022) showed that effects showed up with only 10-100M tokens of training data. Similarly Liu et al. (2022) showed that smaller models with fine tuning may even have better performance than larger models. For these reasons, I selected multiple model sizes, i.e. number of parameters, for Bloom (560M, 1.1B, 3B), XGLM (564M, 1.7B, 2.9B, 4.5B), and Chinese GPT-2 (110M, 3.5B). I tested whether these size of the models correlated with better performance.

2.3 Task

To most closely match the human experimental task, I created a prompt that contained one practice item, which human participants also saw, and then a critical item. Each item consisted of a question about the relative duration estimate or number of subevents for two events (A and B), followed by the two events. In each case, A represented the sentence containing the bare verb and B the sentence containing the reduplicated verb. Examples of the stimuli can be seen in (1) and (2). This task aims to replicated the forced-choice task that human participants completed. The purpose of this was to limit the number of differences between the human and computational experiments.

3 Datasets

3.1 Pilot Behavioral Experiments

To investigate the form-meaning mapping as proposed by MAB, I designed pilot experiments to test whether Mandarin speakers understood reduplicated verbs (VV) as conveying shorter duration than their bare (non-reduplicated; V) counterparts. In these experiments, I asked participants to compare estimated event duration between pairs of V and VV events.

The original human experiments tested whether native Mandarin speakers associated reduplicated forms with either diminishing or increasing semantics. To do this, participants were asked to compare the duration estimate (how long the event took place) or the number of instances (or sub-events) between sentences, where the verb was either bare or reduplicated, and identical in all other respects. Diminishing semantics entails shorter duration and fewer subevents, where increasing semantics entails longer duration and more instances of the event (Arnett, 2022).

The experiments generally found that participants more often associated reduplicated forms with diminishing semantics than with increasing semantics.

3.2 Stimuli

To get the complete prompts, I created a script that took a spreadsheet containing all stimuli from each of the experiments, generated the prompt based on the two events and the question.

Prompt:

(1) 下列哪个活动所需时间更长?
A: 跑马拉松
B: 刷牙
回答: A: 跑马拉松
下列哪个活动所需时间更长?
A:看电视
B:看看电视
回答:

(2) Which of the events below has a longer estimated duration?
A: Run a marathon
B: Brush teeth
Answer: A: Run a marathon
Which of the events below has a longer estimated duration?
A: Watch TV (Bare Verb Condition)
B: Watch-Watch TV (Reduplicated Verb Condition)
Answer:

4 Results

4.1 GPT-2, Bloom, XGLM

All sizes of the GPT-2, Bloom, and XGLM models failed to complete the task. Common errors included answering 'A:' followed by the event that was labeled 'B' in the prompt; answering with both answers; and generating text that was irrelevant and/or incoherent. I tried adding multiple practice items, but was not able to get these models to generate responses, even just the same A or B response for each stimulus.

4.2 GPT-3

For GPT-3, I tested four version of the model, which performed quite differently from each other. Here I present results from experiments 1 and 3 from my previous studies, which have the most simple experimental design.

Visually, it seems clear that text-davinci-003 best models the human results. For experiment 1, davinci and text-davinci-003 capture the general trend of the human results. For experiment 3, which was divided into two sub-experiments, text-davinci-002 best captured the human results for the duration sub-experiment, while all three text-davinci versions captured the general trend for the instances sub-experiment. Overall, the text-davinci models do a fairly good job at capturing the effects seen in the human experiments, especially text-davinci-002 and text-davinci-003.

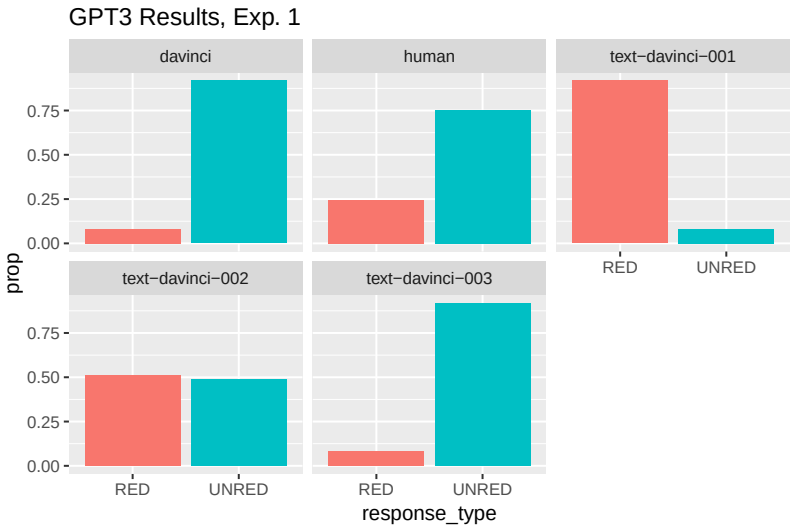

Figure 1: Caption

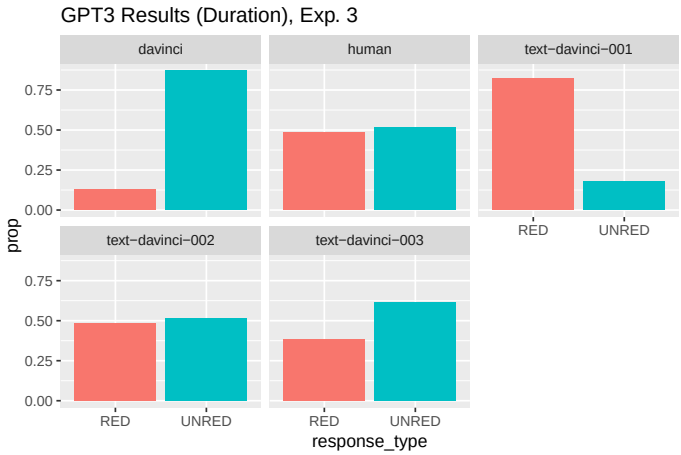

Figure 2: Caption

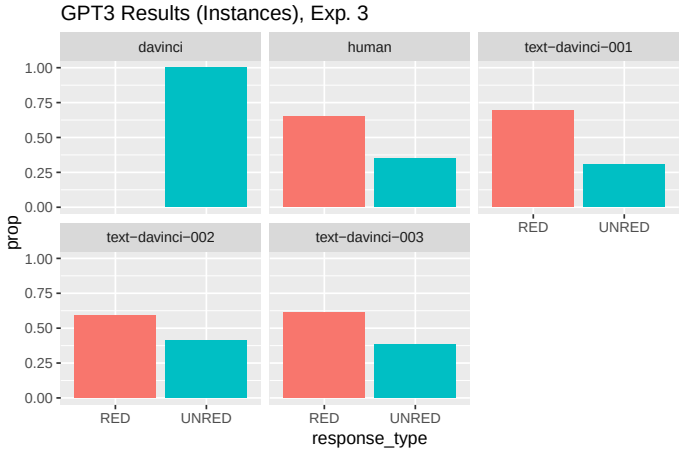

Figure 3: Caption

5 Conclusion

Even though GPT-3 has not been explicitly trained on languages other than English, GPT-3 by far outperformed the other models in this set of experiments. The reason for this may be demonstrated in the difference between the performance of the davinci and text-davinci models. text-davinci represents InstructGPT, which is a version of GPT-3 fine-tuned using reinforcement learning (Ouyang et al., 2022).

As it stands the text-davinci models could potentially serve as pilot subjects for a new psycholinguistic experiment, for example to test whether a stimulus may be problematic for the experiment or to test whether the experimenter might expect a particular effect.

5.1 Future Directions

In the future, to model other human experiments, it may be best to fine-tune a model that has been trained on Chinese explicitly and fine-tune it using methods proposed by Ouyang et al. (2022) or Patel et al. (2022).

References

Arcodia, G. F., Basciano, B., and Melloni, C. (2014). Verbal Reduplication in Sinitic. In *Proceedings of the 8th Décembrettes*, volume 22, pages 15–45, Bordeaux, France.

Arcodia, G. F., Basciano, B., and Melloni, C. (2015). Areal perspectives on total reduplication of verbs in Sinitic. *Studies in Language*, 39(4):836–872.

Arnett, C. (2022). A psycholinguistic approach to form-meaning correspondence in mandarin chinese monosyllabic verbal reduplication. Master's thesis.

Basciano, B. and Melloni, C. (2017). Event delimitation in Mandarin: The case of diminishing reduplication. *Italian Journal of Linguistics*, 29(1):143–166.

BigScience Workshop (2022). Bloom (revision 4ab0472).

Brdar, M. (2013). Adjective reduplication and diagrammatic iconicity. *Sanjari i znanstvenici. Zbornik u čast*, pages 489–514.

Chinese Knowledge and Information Processing (2020). Ckip transformers.

Ek, A., Bernardy, J.-P., and Lappin, S. (2019). Language modeling with syntactic and semantic representation for sentence acceptability predictions. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 76–85.

Facebook (2022). Xglm.

Inkelas, S. and Zoll, C. (2005). *Reduplication: Doubling in morphology*, volume 106 of *Cambridge Studies in Linguistics*. Cambridge University Press.

Lau, J. H., Clark, A., and Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41(5):1202–1241.

Li, B., Zhu, Z., Thomas, G., Rudzicz, F., and Xu, Y. (2022). Neural reality of argument structure constructions. *arXiv preprint arXiv:2202.12246*.

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., Pasunuru, R., Shleifer, S., Koura, P. S., Chaudhary, V., O'Horo, B., Wang, J., Zettlemoyer, L., Kozareva, Z., Diab, M. T., Stoyanov, V., and Li, X. (2021). Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Liu, R., Wei, J., Gu, S. S., Wu, T.-Y., Vosoughi, S., Cui, C., Zhou, D., and Dai, A. M. (2022). Mind's eye: Grounded language model reasoning through simulation. *arXiv preprint arXiv:2210.05359*.

Marantz, A. (1982). Re reduplication. *Linguistic inquiry*, 13(3):435–482.

Marvin, R. and Linzen, T. (2018). Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

Melloni, C. and Basciano, B. (2018). Reduplication across boundaries: The case of Mandarin. In *The lexeme in descriptive and theoretical morphology*, volume 4 of *Empirically Oriented Theoretical Morphology and Syntax*, pages 325–363. Language Science Press.

Michaelov, J. A. and Bergen, B. K. (2020). How well does surprisal explain n400 amplitude under different experimental conditions? *arXiv preprint arXiv:2010.04844*.

Misra, K., Ettinger, A., and Rayz, J. T. (2020). Exploring bert's sensitivity to lexical cues using tests from semantic priming. *arXiv preprint arXiv:2010.03010*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Patel, A., Li, B., Rasooli, M. S., Constant, N., Raffel, C., and Callison-Burch, C. (2022). Bidirectional language models are also few-shot learners.

Prasad, G., Van Schijndel, M., and Linzen, T. (2019). Using priming to uncover the organization of syntactic representations in neural language models. *arXiv preprint arXiv:1909.10579*.

Sinclair, A., Jumelet, J., Zuidema, W., and Fernández, R. (2022). Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Warstadt, A., Singh, A., and Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.