Predicting Reference: What do Language Models Learn about Discourse Models?

Shiva Upadhye
Department of Cognitive Science
UC San Diego
supadhye@ucsd.edu

Leon Bergen
Department of Linguistics
UC San Diego
lbergen@ucsd.edu

Andrew Kehler
Department of Linguistics
UC San Diego
akehler@ucsd.edu

Abstract

Whereas there is a growing literature that probes neural language models to assess the degree to which they have latently acquired grammatical knowledge, little if any research has investigated their acquisition of discourse modeling ability. We address this question by drawing on a rich psycholinguistic literature that has established how different contexts affect referential biases concerning who is likely to be referred to next. The results reveal that, for the most part, the prediction behavior of neural language models does not resemble that of human language users.

1 Introduction

The impressive power of deep learning based language models has inspired a new line of computational psycholinguistics research that examines the extent to which linguistic knowledge lies latent within their distributed networks. This work has primarily focused on linguistic phenomena that syntactic theory tells us requires syntactic knowledge to capture, with mixed results (Linzen et al. 2016; Lau et al. 2017; Goldberg 2019; Warstadt et al. 2019; inter alia). This paper asks a new question: to what extent do these language models capture the linguistic knowledge required to perform *discourse* modeling?

We are unaware of any work that has addressed this question directly. Perhaps the closest research has centered on the Winograd Schema Challenge (WSC) (Levesque et al., 2012), which evaluates the ability of systems to employ world knowledge to interpret ambiguous pronouns in minimal pairs that resemble Winograd's famous example (1).

(1) The city councilmen refused the demonstrators a permit because

a. they feared violence. [they = city council]

b. they advocated violence. [they = demonstrators]

However, WSC is essentially a 'fill in the blank' problem-solving task, and doesn't evaluate the extent to which systems display humanlike ability to model discourse in an online, incremental fashion. We instead take our inspiration from psycholinguistic work that has focused on this question. For instance, the Bayesian Model of pronoun interpretation (Kehler et al., 2008; Kehler and Rohde, 2013) posits that comprehenders resolve the meaning of a pronoun via Bayesian principles by combining their estimates of the speaker's production biases (the LIKELIHOOD term) with their top-down expectations about which entities are likely to be mentioned next (the PRIOR term, which we refer to as the NEXT-MENTION BIAS). Kehler and Rohde (2013) demonstrate that an array of semantic biases (e.g., verb semantics) and pragmatic biases (e.g., coherence relations) that have been claimed to influence pronoun interpretation directly actually do so only indirectly, by conditioning the prior.

The role of the prior in the Bayesian Model is directly analogous to its role in Bayesian approaches to tasks such as speech recognition and machine translation, where a language model provides the prior probabilities. We argue that the ability to capture the influence of context on next-mention biases is thus a particularly appropriate task for evaluating the extent to which language models capture discourse modeling knowledge. Our focus will be on effects of verb semantics that the psycholinguistic literature has shown to influence next-mention biases. These studies have used a PASSAGE COMPLETION paradigm, in which experimental participants are presented with context clauses followed by either a full stop (2a) or a conjunction (2b-c), and asked to complete the passage with the first follow-on sentence that comes to mind.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 977–982,
November 16–20, 2020.

(2) a. John impresses Mary. _____
b. John impresses Mary because _____
c. John impresses Mary, and as a result ___

Analysis of the completions yields estimates of next-mention biases and of referential form production. In the task described in §3, we will probe the next-mention biases produced by two language models in different contexts that we describe now.

2 Comparisons and Predictions

If neural language models latently acquire discourse modeling knowledge, they should be able to distinguish between contexts that are superficially similar but which are known from psychological research to yield significant effects on next-mention biases. We focus on three such contrasts.

Implicit Causality Verbs The first comparison is between two kinds of so-called IMPLICIT CAUSALITY (IC) verb, exemplified in (3a-b).

(3) a. John aggravated Mary. [IC1]
b. John praised Mary. [IC2]

Sentences with IC verbs generate an expectation that the follow-on sentence will participate in an Explanation coherence relation, in which the second sentence provides a cause or reason for the eventuality described by the first (Kehler et al., 2008). However, the two types differ in which event participant causality is attributed to. IC1 verbs (3a) have been experimentally shown to generate a strong expectation that the preceding subject will be mentioned next in the follow-on sentence—we heard that John is aggravating, and we now expect to hear why (Garvey and Caramazza 1974; Caramazza et al. 1977; Brown and Fish 1983; Terry Kit-fong Au 1986; McKoon et al. 1993; Koornneef and van Berkum 2006; Kehler et al. 2008; inter alia). IC2 verbs (3b), on the other hand, have been shown to generate a strong expectation that the preceding object will be mentioned next in the follow-on sentence—we heard that Mary received praise, and we now expect to hear why. We can then ask: do IC1 and IC2 verbs generate different expectations in language models for next mention in otherwise identical contexts?

There are also subsidiary predictions regarding the use of connective prompts as in (2b-c). For both types of IC verbs, *because* prompts strengthen their biases, since virtually 100% of the continuations will now be Explanations rather than 60% as found in full stop prompt conditions (Kehler et al., 2008). So we expect to see a higher probability of next-mention of the subject with *because* prompts for IC1 verbs, and likewise for objects for IC2 verbs. Both types of IC verb, however, are known to have a strong bias to the object in Result coherence relations (Stewart et al., 1998; Kehler et al., 2008)—in which the follow-on describes an effect rather than a cause—which are enforced by the *and as a result* prompt. For IC1 verbs, therefore, we should see a strong shift toward the object with *and as a result* prompts compared to full stop prompts. To summarize the predictions:

1a. IC1 contexts with full stop prompts should display a stronger next-mention bias to the subject compared to IC2 contexts.

1b. Contexts with *because* prompts should strengthen the next-mention bias associated with each type of verb compared to full stops.

1c. *And as a result* prompts in IC1 contexts should result in a greater next-mention bias toward the object compared to full stops.

Motion vs. Transfer of Possession Verbs The second comparison is between Motion (4a) and Transfer-of-Possession (ToP) verbs (4b).

(4) a. The man jogged to the woman. [Motion]
b. The man handed a gift to the woman. [ToP]

These sentence types are superficially similar: they each have a grammatical subject that functions as a thematic Agent/Source, and a grammatical object-of-preposition that functions as a thematic Goal. However, they are known to yield very different next-mention biases. Specifically, previous studies have revealed that whereas motion verbs have a strong next-mention bias toward the previous subject (e.g., 84.4% in a study run by Stevenson et al. (1994)), ToP contexts give rise to a distribution that's closer to 50/50 (51.0%). The reason is that the Goal in ToP sentences functions not only as a location but a recipient as well, leading to an expectation that we'll next hear about what the recipient did with the object of transfer, which counteracts the typical subject bias. We thus expect to see a much stronger next-mention bias toward the subject for Motion contexts as compared to ToP contexts, despite their superficially similar properties. Further, we expect a large effect of the connective

conditions: previous work (Stevenson et al., 1994; Kehler et al., 2008) has shown Explanations to be strongly biased to the Source, and Result continuations to be strongly biased to the Goal for ToP contexts.[1] To summarize the predictions:

2a. Motion contexts with full stop prompts should display a stronger next-mention bias to the subject compared to ToP contexts.

2b. ToP contexts with *because* prompts should yield a stronger bias toward the subject compared to full stop prompts.

2c. ToP contexts with *and as a result* prompts should yield a stronger bias to the object compared to full stop prompts.

Aspectual Marking with Transfer of Possession Verbs The final comparison varies aspectual marking rather than the semantic class of the verb. Kehler et al. (2008) compared ToP contexts in the perfective such as (4b) with otherwise identical sentences in the imperfective (5):

(5) The man was handing a gift to the woman.

Following Stevenson et al. (1994), Kehler et al. conjectured that ToP verbs have a special property in that the prominence of the event participants depends on what component of event structure is being focused on. Specifically, the imperfective focuses the hearer's attention on the ongoing development of the event, where the agent of the event is most prominent. The perfective (4b), on the other hand, focuses the hearer's attention on the end state of the event, where the recipient becomes prominent. Kehler et al. therefore predicted that imperfective contexts would lead to a greater referential bias to the agent than perfective contexts, which is precisely what they found (80% vs. 57%). This gives rise to the following prediction:

3. Imperfective ToP contexts should display a stronger next-mention bias to the subject compared to perfective ToP contexts.

3 Experimental Setup

We evaluated two state-of-the-art, pre-trained autoregressive language models (LMs): GPT-2 large (Radford et al., 2018) and Transformer-XL (Dai et al., 2019).[2] The experiments were conducted in a zero-shot setting, and the task of generating continuations was reformulated to a next-word prediction task. Prior to tokenization, the input stimulus was prepended with a token indicating the beginning of the sentence. Additionally, the inputs for Transformer-XL were prepended with a padding text to account for the shorter stimulus length.[3]

Miss Smith ___ Mr. Smith	Mary ___ John
The woman ___ the man	Alice ___ Bob
The actress ___ the actor	The girl ___ the boy
Mrs. Taylor ___ Mr. Williams	Emma ___ David
The princess ___ the prince	Sarah ___ Robert
Mrs. Williams ___ Mr. Taylor	Emily ___ Paul

Table 1: Context Sentence Frames

To capture the diversity of ways in which event participants can be mentioned in the context sentence, the twelve frames shown in Table 1 were used. In order to balance for the effects of gender (Zhao et al., 2018; Bordia and Bowman, 2019), each frame was used again with the order of the event participants reversed, for a total of 24 frames. 20 IC1 verbs, 20 IC2 verbs, 18 Motion verbs, and 18 ToP verbal complexes (in both perfective and imperfective variants) were each run in the full stop prompt, *because* prompt, and *and as a result* prompt conditions, in each of the 24 frames.[4]

After presenting a pairing of a context sentence and prompt, we compute the (normalized) conditional probabilities of *He* and *She* in the full stop prompt condition and their lowercase equivalents for the connective prompt conditions. The average biases to the subject are computed for each verb over the sentence frames, which are in turn averaged to compute the overall subject bias for each context type. The latter averages are reported with 95% confidence intervals in the tables below.

4 Results

Implicit Causality Comparison The next-mention biases toward the subject produced by each system in the IC verb conditions are shown in Tables 2 and 3.

[1] Unfortunately, Stevenson et al. (1994) left Motion contexts out of their experiment that examined the role of connectives. We thus have no data to compare to for these conditions.

[2] We considered also evaluating BERT on this task but decided that it was unsuitable. BERT performs masked language modeling, conditioned on both left and right contexts. The current experiments use only the left context, and hence BERT would need to be queried in a non-natural setting.

[3] For padding text see: https://tinyurl.com/y9kjuj5q.

[4] The actual verbs used and other information necessary for reproducibility of results has been placed at https://github.com/shiva-upadhye/predicting-reference.

Prompt	Transformer-XL	GPT-2
full stop	.51 ± .01	.59 ± .02
because	.61 ± .03	.63 ± .02
and as a result	.43 ± .02	.31 ± .02

Table 2: Subject next-mention bias for IC1 contexts

Prompt	Transformer-XL	GPT-2
full stop	.51 ± .02	.66 ± .02
because	.45 ± .02	.42 ± .05
and as a result	.50 ± .05	.47 ± .07

Table 3: Subject next-mention bias for IC2 contexts

Our first question (Prediction 1a) is whether the LMs would display a greater next-mention bias toward the preceding subject in IC1 contexts than IC2 contexts. The answer is no: As can be seen in the first rows of Tables 2 and 3, the biases across conditions for Transformer-XL are identical (.51) and the difference witnessed for GPT-2 goes in the wrong direction (.59 vs. .66). These results therefore do not align with the more polar biases for IC contexts that the psycholinguistic literature has revealed in human studies.

The second question (Prediction 1b) is whether the occurrence of *because* at the end of the prompt—which for human language users shifts discourse coherence expectations toward Explanation continuations—strengthens the respective IC biases. This prediction receives only limited support: The results in Table 2 reveal increased biases toward the subject compared to the full stop condition for IC1 verbs, and those in Table 3 reveal similar decreases for IC2 verbs. However, only GPT-2 in the IC2 condition yielded an effect of the magnitude that human language studies might lead us to expect.[5]

The final question (Prediction 1c) is whether the occurrence of *and as a result* at the end of the prompt—which for human language users shifts discourse coherence expectations toward Result continuations—generates a stronger bias toward the preceding object compared to the free prompt baseline in IC1 contexts. This prediction was confirmed for GPT-2, where the connective prompt reduced the bias to the subject by .28. Whereas Transformer-XL witnessed a lower bias in this condition as well, the effect was smaller (.08).

[5]For instance, Kehler et al. (2008) found subject biases of 85% and 60% for IC1 verbs in the *because* and full stop prompt conditions respectively.

To sum, both models failed to yield the hypothesized effect of verb type in the full stop condition. However, there was some degree of sensitivity to the occurrence of a connective, with GPT-2 in particular displaying a strong numerical difference compared to the free prompt baseline in all but the IC1/*because* condition.

Motion vs. ToP Verb Comparison The next-mention biases toward the subject produced by each system in the Motion and ToP context conditions are shown in Tables 4 and 5.

Prompt	Transformer-XL	GPT-2
full stop	.57 ± .01	.63 ± .01
because	.61 ± .02	.65 ± .01
and as a result	.54 ± .02	.47 ± .02

Table 4: Subject next-mention bias for Motion verbs

Prompt	Transformer-XL	GPT-2
full stop	.52 ± .01	.54 ± .03
because	.53 ± .03	.53 ± .03
and as a result	.47 ± .04	.26 ± .03

Table 5: Subject next-mention bias for ToP verbs (perfective)

Our first question (Prediction 2a) asked whether the LMs would display a greater next-mention bias toward the preceding subject in Motion contexts than ToP contexts in the full stop condition. The answer is mostly no: Whereas there is a small numerical difference for each system in the right direction, it is far from what the results of experimental studies would predict. In particular, whereas the bias found for ToP verbs is aligned with established experimental results, the expected strong subject bias for Motion verbs did not materialize.

The second and third questions (Predictions 2b and 2c) asked about the effect of connectives in the ToP condition, whereby *because* and *and as a result* prompts should pull expectations toward the subject and object compared to the full stop prompt baselines respectively. As with IC verbs, no strong effect was witnessed for Transformer-XL, whereas GPT-2 did show a strong shift in the predicted direction for *and as a result* prompts. However, no appreciable effect was seen for GPT-2 in the *because* prompt condition.

Aspectual Marking in ToP Verbs Comparison Our final question (Prediction 3) probes the poten-

tial effects of aspectual marking on next-mention biases, in particular whether imperfective ToP contexts will yield a stronger next-mention bias to the subject compared to perfective ToP contexts. The results for perfective and imperfective ToP contexts are shown in Tables 5 and 6 respectively.

Prompt	Transformer-XL	GPT-2
full stop	.57 ± .01	.62 ± .02
because	.56 ± .03	.57 ± .02
and as a result	.63 ± .03	.45 ± .03

Table 6: Subject next-mention bias for ToP verbs (imperfective)

Prediction 3 was mostly disconfirmed: There is only a modest difference between ToP contexts using the perfective and imperfective aspect in the full stop prompt condition. Interestingly, however, the predicted effect did exist for both systems in the *and as a result* condition. It is not clear to us why the effect would be limited to only this condition.

5 Conclusions

We set out to evaluate the extent to which neural LMs latently acquire the discourse modeling capability necessary to perform a particular type of incremental processing that human language users do: The ability to predict what entities are most likely to be mentioned next. We examined three context pairs with superficially similar linguistic properties that the experimental literature has shown to result in divergent next-mention biases, both with and without connectives.

The results were mostly, but not entirely, negative. On the one hand, we found no compelling evidence that the LMs are sensitive to any of the three manipulations within the verbal complex in the context sentence. On the other hand, one could argue for preliminary support for the claim that one of the LMs—GPT-2—is sensitive to the occurence of the two connectives examined here. Future work will be required to assess the extent to which these effects do in fact reflect the acquisition of a latent form of discourse modeling ability.

Our conclusions, of course, remain preliminary in a number of respects. First, we have analyzed the behavior of only two systems. Since each system can be said to stand proxy for a single experimental participant, these results could be argued to be less robust than human language studies, which typically utilize several dozen participants. Whereas this limitation is shared with previous work that probes LMs for inherently acquired syntactic knowledge, the robustness of the findings would be enhanced by examining a broader range of systems and/or system configurations so as to better capture the kinds of variation found among groups of human participants.

Second, we have focused here on broad contrasts between context types that have been studied in the psycholinguistic literature. Although the stimuli employed were modeled after those used in experimental studies, to improve the robustness of the findings we felt it necessary to compute means over a variety of sentence frames (Table 1), so that any idiosyncrasies of particular frames that are independent of the manipulation under scrutiny wouldn't unduly (and undetectably) drive the results. This improves the robustness of our results in terms of items—whereas participants in psycholinguistic studies typically see only one example sentence for each verb, the LMs here saw 24—it also means that no lab data exists for the exact stimuli used here. Since an experiment that collects data on this scale would require a substantial annotation effort, a more careful comparison of this sort must be left for future work.

Third, there are many variations of the studies presented here that could be attempted. Examples would include variants that employ longer and more realistic contexts. In this initial investigation we focused on single-sentence contexts so as to hew as closely as possible to previous experimental work. We hope that this short paper will inspire further research that takes next steps in this and a variety of other directions.

Finally, we want to be clear that we do not claim that the two LMs examined have in any sense 'failed' at this task—they were obviously not trained for this purpose. Our goal instead was to pose the novel question of to what extent discourse knowledge of the sort examined here may exist latently in the models. That having been said, we consider the identification of alternative language model architectures that are capable of capturing the requisite discourse modeling capability for this task to be an interesting challenge problem for future work.

Acknowledgments

We thank three anonymous reviewers and our area chair for helpful feedback.

References

Terry Kit-fong Au. 1986. A verb is worth a thousand words: The causes and consequences of interpersonal events implicit in language. *Journal of Memory and Language*, 25:104–122.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Roger Brown and Deborah Fish. 1983. The psychological causality implicit in language. *Cognition*, 14:237–273.

Alfonzo Caramazza, Ellen Grober, Catherine Garvey, and Jack Yates. 1977. Comprehension of anaphoric pronouns. *Journal of Verbal Learning and Verbal Behaviour*, 16:601–609.

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.

Catherine Garvey and Alfonzo Caramazza. 1974. Implicit causality in verbs. *Linguistic Inquiry*, 5:549–564.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Andrew Kehler, Laura Kertz, Hannah Rohde, and Jeffrey L. Elman. 2008. Coherence and coreference revisited. *Journal of Semantics*, 25(1):1–44.

Andrew Kehler and Hannah Rohde. 2013. A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, 39(1-2):1–37.

Arnout W. Koornneef and Jos J. A. van Berkum. 2006. On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye-tracking. *Journal of Memory and Language*, 54:445–465.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 552–561, University of Rome.

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.

Gail McKoon, Steven B. Greene, and Roger Ratcliff. 1993. Discourse models, pronoun resolution, and the implicit causality of verbs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18:266–283.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

Rosemary J. Stevenson, Rosalind A. Crawley, and David Kleinman. 1994. Thematic roles, focus, and the representation of events. *Language and Cognitive Processes*, 9:519–548.

Andrew J. Stewart, Martin J. Pickering, and Anthony J. Sanford. 1998. Implicit consequentiality. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, pages 1031–1036.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.