Do Neural Language Models Compose Concepts the Way Humans Can?

Amilleah Rodriguez[1,2], **Shaonan Wang**[1,2], **Liina Pylkkanen**[1,2]
[1]Department of Linguistics, New York University
[2]Department of Psychology, New York University
{amilleah.rodriguez, shaonan.wang, liina.pylkkanen}@nyu.edu

Abstract

While compositional interpretation is the core of language understanding, humans also derive meaning via inference. For example, while the phrase “the blue hat” introduces a blue hat into the discourse via the direct composition of “blue” and “hat,” the same discourse entity is introduced by the phrase “the blue color of this hat” despite the absence of any local composition between “blue” and “hat.” Instead, we infer that if the color is blue and it belongs to the hat, the hat must be blue. We tested the performance of neural language models and humans on such inferentially driven conceptual compositions, eliciting probability estimates for a noun in a syntactically composing phrase, "This blue hat", following contexts that had introduced the conceptual combinations of those nouns and adjectives either syntactically or inferentially. Surprisingly, our findings reveal significant disparities between the performance of neural language models and human judgments. Among the eight models evaluated, RoBERTa, BERT-large, and GPT-2 exhibited the closest resemblance to human responses, while other models faced challenges in accurately identifying compositions in the provided contexts. Our study reveals that language models and humans may rely on different approaches to represent and compose lexical items across sentence structure. *All data and code are accessible at* `https://github.com/wangshaonan/BlueHat`.

Keywords: Neural Language Models, Composition, Inference, Dataset Construction

1. Introduction

Language comprehension involves combining word meanings according to the structure of a sentence and yet, it also encompasses a wide range of inferential processing. Neuroscience research has revealed a uniform brain basis for conceptual combinations that align with syntax and those that result through inference (Parrish et al., 2022). Building on this finding, we tested human and language model performance on inferentially vs. syntactically arising conceptual combinations. Can neural language models, proficient in diverse language tasks (Han et al., 2021) and mirroring patterns of human brain activity (Sun et al., 2019; Wang et al., 2020; Caucheteux and King, 2022) create combined concepts even when the combinations are not obvious from the syntax? To investigate this phenomenon in neural language models, we introduce a novel dataset introducing pairs of sentences matched in syntactic structure but which vary in whether they introduce a conceptual combination of an adjective and noun or not.

1. *The blue color of this hat is lovely. This blue hat ..*
2. *The blue lamp near this hat is lovely. This blue hat ..*

Despite maintaining the same syntactic distance between "blue" and "hat" in both context sentences, humans only obtain a "blue hat" interpretation in the first case. Therefore, only in the first case is a subsequent reference to a blue hat natural. This study tests whether the same information is available for language models: can they represent combined concepts even when the two elements do not syntactically merge and the combination arises via inference?

Previous research has explored neural language models’ syntactic and semantic capabilities, investigating elements like filler-gap relationships (Wilcox et al., 2018; Linzen and Baroni, 2021), subject-verb agreements (Linzen et al., 2016; Jawahar et al., 2019), anaphor binding (Hu et al., 2020), non-syntactic factors like politeness effects (Lee and Wang, 2023), semantic prowess like quality of phrase representations (Yu and Ettinger, 2020; Garcia et al., 2021), and the capacity to recognize semantic roles and possess event knowledge (Ettinger, 2020; Pavlick, 2022; Kauf et al., 2022). The current study adds to this body of work by asking whether humans and language models diverge in their ability to combine concepts in the absence of a local syntactic relationship.

2. Methodology

In our study, we provided a context sentence using color and object descriptors, for example, “The blue color of this hat is lovely.” This was followed by a probe expression like “This blue hat...” When the initial sentence combined the concepts of color and object as in the probe, the term ‘hat’ is a natu-

LREC-COLING 2024, pages 5309–5314

ral and predictable continuation after 'blue' in the probe sentence. In contrast, after a context sentence such as "The blue lamp near this hat is lovely," which does not introduce the combined concept of the probe expression, the predictability of the noun "hat" is effectively zero after blue. Thus, human participants who effortlessly discern the long-distance semantic relationship between 'blue' and 'hat' in the context sentence are expected to assign a higher score to the naturalness of 'hat' as the next word after 'blue' in the probe when the context sentence introduced the combined concept of a blue hat via inference. Similarly, a language model that can effectively capture the long-distance semantic connections in our context sentence will assign probabilities based on whether or not the context introduces the combined concept of a blue hat.

Altogether, our context sentences introduced eight distinct compositional contexts for the probe expression, as shown in Section 3.1. Human judgments were log-transformed naturalness ratings of the noun in the probe expression. To assess whether the language model inferred a 'blue hat' from each preceding context, we evaluate **surprisal** (Hale, 2016) as the log-transform of output probabilities at the word 'hat' in our probe expression. Additionally, since our context manipulation also may affect how predictable the determiner 'This' is at the beginning of the probe, we subtracted the probability of 'This' from the probability of 'hat' in our analysis. This adjustment helps us account for how each context influences the interpretation of the probe phrase. We then contrasted surprisal across minimal pairs of combinatory and non-combinatory contexts within matched target items and syntactic distance. We used a binary measure to categorize language model performance across these minimal pairs, correctly determining 'hat' in the probe as less surprising when the context introduced a blue hat, than when the context did not. We used this binary measure to calculate percent **accuracy** across each condition of combinatory and non-combinatory pairs as shown in Table 1.

To further investigate how language models represented and used these semantic relationships over linear and syntactic distances, we also examined the **information retention** of the target adjective on the noun (or noun on the adjective, depending on word order) within our context sentences. We computed the cosine similarity between the adjective and noun at each word position within the context sentence, with the assumption that features relevant for composition would be maintained across syntactic distance. The results for our best performing model, GPT-2, are displayed in Figure 2., with the rest of the language models results in Figure 3.

3. Experimental Setting and Dataset

3.1. BlueHat dataset

We followed established psycholinguistic methods to minimize the effects of word frequency and sentence structure variations using a controlled vocabulary of five color adjectives and five nouns sorted into sentence templates. We included sets of prepositions, verbs, adverbs, intensifiers, and color descriptors to avoid the repetitive use of a small vocabulary. The frequencies of these elements were balanced within each template item. This method produced 100 sentence pairs across eight distinct conditions (800 context sentences and corresponding probes).

1. **Local(ADJ)**
   *The color of the **blue hats** is lovely.* **(combinatory)**
   *The lamps are **blue**, **hats** red, and socks gray.* **(non-combinatory)**
2. **Local(NOUN)**
   *You'll see that the **hat** is **blue** in color.* **(combinatory)**
   *You will see a **hat**, a **blue** lamp and socks.* **(non-combinatory)**
3. **Non-local(ADJ)**
   *The **blue** color of this **hat** is lovely.* **(combinatory)**
   *The **blue** lamp near this **hat** is lovely.* **(non-combinatory)**
4. **Non-local(NOUN)**
   *The **hat** is surely a lovely **blue** color.* **(combinatory)**
   *The **hat** is near a lovely **blue** lamp.* **(non-combinatory)**

In these examples, Non-local refers to the inferential, long-distance context the target adjective and noun appear in, while Local indicates local contexts of syntactic composition. We use (ADJ) or (NOUN) to indicate which of the target words, the adjective or noun, appears first within the sentence. Combination, either combinatory or non-combinatory refers to whether the target adjective and noun can be composed within the sentence.

3.2. Human behavioral experiment

We conducted a behavioral experiment involving 40 participants recruited from Prolific (Palan and Schitter, 2018). These participants were tasked with evaluating the naturalness of 120 sentences, comprising 15 pairs of context sentences and corresponding probes randomly selected from a pool of 100 sentence pairs. Participants utilized a 5-point scale, ranging from 1 (Very Unnatural) to 5

Model	Overall	Local (ADJ)	Local (NOUN)	Non-local (ADJ)	Non-local (NOUN)
Humans	94.3	95.4	97.1	89.8	95.0
GPT	48.25	66.0	44.0	41.3	50.7
GPT-2	71.75	96.0	100.0	86.0	40.0
GPT-2-large	22.0	72.0	1.67	16.0	19.3
BERT	43.75	20.0	16.0	68.7	36.0
BERT-large	58.5	68.0	60.0	44.0	62.7
DistilBERT	34.75	62.0	30.0	20.7	41.3
RoBERTa	58.0	100.0	94.0	19.3	70.7
XLNet	26.75	10.0	2.0	36.0	39.3

Table 1: Human and Language Model Accuracy across Word Order and Distance.

(Very Natural), to rate the degree to which the probe expression felt like a natural continuation of the context sentence. The task required participants to assess the coherence of these sentence pairs. The entire experiment lasted 25 minutes, and participants were compensated with $12 per hour for their time. To ensure the quality of responses and maintain participant engagement, we incorporated response checkpoints. Participants who scored below 80% accuracy on these response checks were excluded from the subsequent analyses.

3.3. Language models

The present study aims to identify strategies neural language models 1 use to represent and compose meaning in adjective-noun phrases. We tested three distinct model groups: 1) BERT Family: We utilized BERT, BERT-large, distillBERT, and RoBERTa that employ bidirectional learning. Notably, RoBERTa's singular objective function sets it apart from other BERT variants. 2) GPT Family: This group includes autoregressive training models like GPT, GPT-2, and GPT-2-large, valuable for sequential text generation tasks requiring creativity. 3)XLNet: Positioned at the crossroads of BERT and GPT families, XLNet combines bidirectional architecture and autoregressive training methods. GPT-2 and RoBERTa exhibit surprisal that most closely resembles human naturalness ratings collected in our experiment. Their performance suggests they may hold more flexible representations to extend the meaning of a sentence to a novel combination. The other models we evaluated scored below chance accuracy comparing combinatory and noncombinatory sentence minimal pairs. [1].

4. Results

Table 1 reveals a striking disparity: all language models underperform in comparison to human judg-

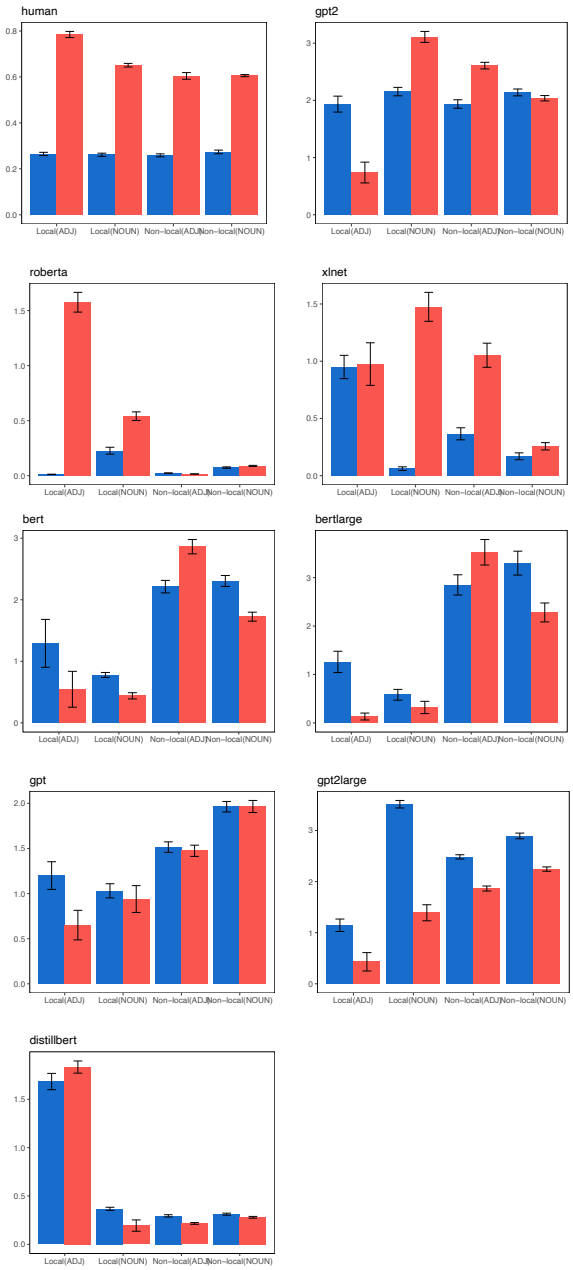

Figure 1: Mean human naturalness judgments (top-left) and language model surprisal to the probe sentence for combinatory (blue) and non-combinatory (red) contexts.

[1]All models used in this study are sourced from OpenAI's pre-trained models in the Transformers library https://huggingface.co/transformers/v3.3.1/pretrained_models.html

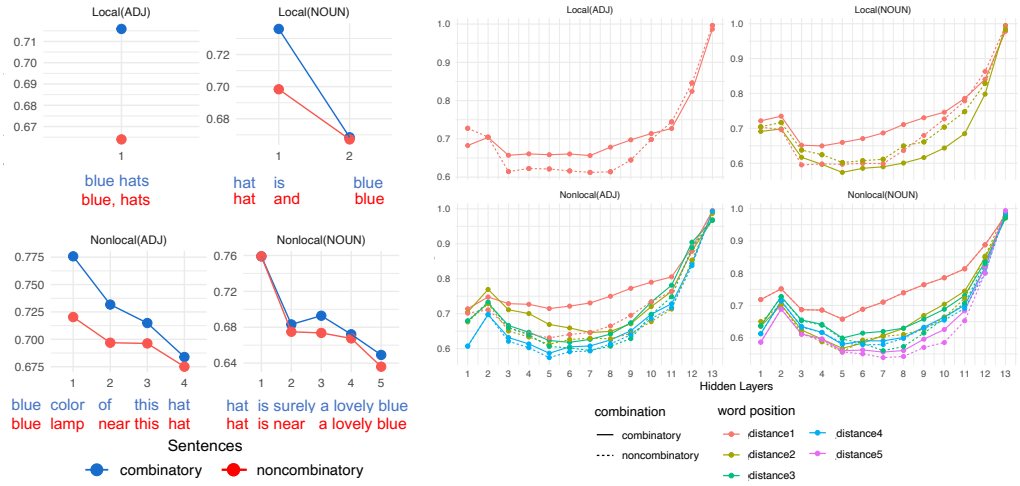

Figure 2: (GPT-2) Cosine similarity of target words across sentence (left) and target words across sentence at each hidden layer (right) for combinatory (blue) and non-combinatory (red) contexts.

ments. The overall accuracy for human participants was 94.3%, with higher accuracies obtained on the Local contexts than Non-local. This indicates participants' ability to reliably identify when sentences were forming the intended adjective-noun structure but reflects a decrease in the naturalness of the probe's minimal construction when the items composed through inference.

Among the language models, GPT-2 exhibited the highest overall accuracy at 72%. RoBERTa performed well in our Local control contexts and showed above-chance performance on the Non-local (NOUN) condition. However, it struggled with an accuracy of only 19.3% across the Non-local (ADJ) sentences, indicating difficulty in extracting the concept of a blue hat from the phrase "blue color of this hat." Additionally, while other tested models demonstrated above-chance accuracy on minimal pairs of sentences, their low accuracy across our Local control conditions suggests a failure to recognize compositional structure, even when words occurred in linear order. This disparity highlights a key difference between human understanding and current language models' capabilities in comprehending complex linguistic structures.

Figure 1 displays naturalness ratings for humans and surprisal for the language models we tested. Humans consistently identify the probe in non-combinatory contexts as less natural than in combinatory contexts across distances and order. For instance, 'the blue hat' in a non-combinatory context might follow a sentence like 'the blue lamp near this hat is lovely,' making it unexpected to directly mention a 'blue hat.' However, the models' behavior differed from that of humans. GPT-2 assigned higher surprisal to combinatory conditions in Local(ADJ) and more similar surprisal values for the Non-local(NOUN) conditions. Surprisal results for the other models revealed that BERT, GPT, and GPT-2-large failed to correctly assign higher probabilities to our combinatory, Local control sentences, indicating that they were unable to detect when composition was available. While mean surprisal indicates XLNet performed better on these tasks overall, its accuracy indicates poor performance between minimal pairs of sentences.

For GPT-2, an autoregressive model, this difficulty in Non-local conditions indicates that it may have difficulty in maintaining long-range dependencies across the input, leading to challenges in accurately identifying and understanding compositional structures that span multiple words or phrases. RoBERTa, a transformer-based model, may have faced challenges in effectively using contextual information from distant parts of the input sequence to accurately identify compositional structures in Non-local contexts and extend them to novel combinations.

5. Information Retention

Information retention across context sentence An additional analysis was designed to investigate how language models build complex representations of their text input. If a word's representation is used later on in a sentence, such as when a word conceptually combines with another word across our Non-local contexts, that representation should be maintained or accessed at later points in the sentence. We measured the cosine similarity between the target noun and target adjective in the context sentence at every word between them. Figure 2 (left) shows cosine similarity scores from GPT-2 across our context sentences. The cosine similarity score between the target noun and the target adjective at each point in the context sentence indi-

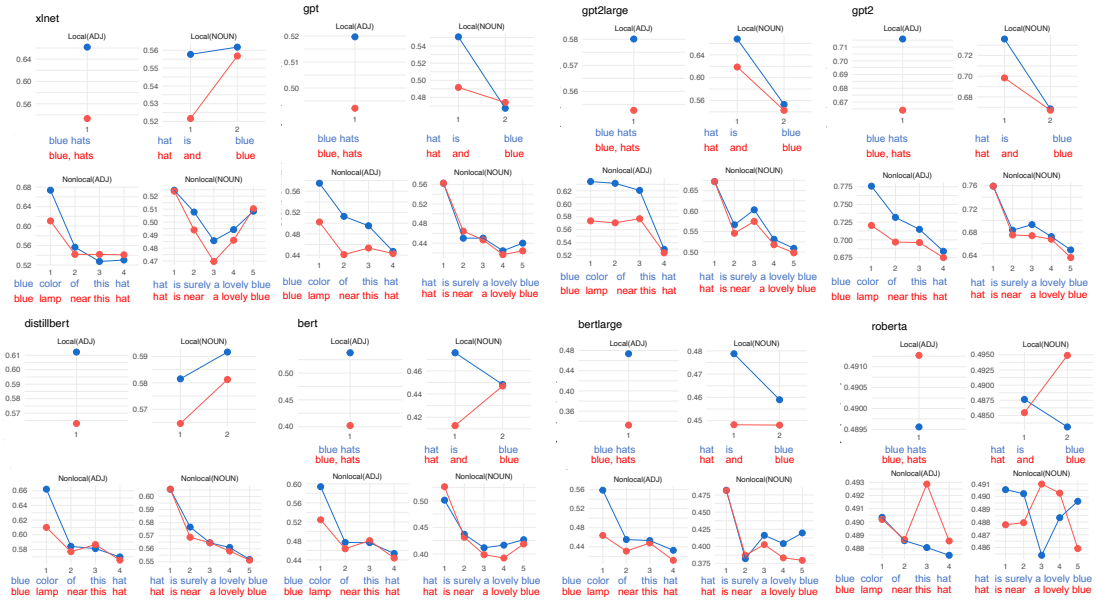

Figure 3: Language model results. Cosine similarity of target words across sentence for combinatory (blue) and non-combinatory (red) contexts.

cates how closely related the current word is to the representation of the context's noun. If the cosine similarity is high, it suggests that information relevant for the representation of the composing items, is being maintained or used at the current word. This analysis explains the decrease in accuracy within the Non-local conditions. The convergence of cosine similarities for GPT-2 as the sentence progresses suggests a difficulty in maintaining representations relevant for conceptual combination across extended spans of text, potentially contributing to its challenges in associating adjective-noun pairs in the probe sentence. Despite this, it still correctly has a higher cosine similarity between the two target items when they can compose. This shows an understanding of the context sentence, even if the model has difficulty evaluating the probe.

Information retention across hidden layers For GPT-2, we were also interested in understanding how the model maintained and used compositional information in our experiment. To do this, we took the same cosine similarity measure, between the adjective and noun at each word position, but output these results at each hidden layer. This allows us to identify when the model distinguishes between composing and non-composing sentences, even if that information is lost as the model's representations become more complex. Figure 2 (right) shows the cosine similarity between the representations of the word "hat" from the probe and the word "blue" from the context across different sentence structures (denoted by distances 1 to 5) and across the 13 layers of GPT-2. For both local and non-local conditions, the combinatory context generally results in higher cosine similarity, suggesting GPT-2 is able to discern when "hat" is a predictable continuation of "blue". The increase in cosine similarity in higher layers suggests that as the network's deeper layers build more complex representations of their input, the distinction between word-level items is lost. This suggests that using the middle or earlier layers may be helpful in further investigations of the conceptual capabilities of neural language models.

6. Conclusion

This paper contributes to the ongoing efforts to deepen our understanding of the intricate linguistic capabilities exhibited by neural language models. We introduced a novel test that examines whether language models can understand conceptual combinations even when they arise via inference. Our findings show that language models generally did not perform similarly to humans but found that GPT-2 was the most accurate out of the language models we tested. GPT-2's correct assignment of higher cosine similarity across sentence structure, but higher surprisal for combinatory contexts reveals that it has the ability to identify when conceptual combinations are introduced, but it lacks the ability to maintain that information in our long-distance conditions. Each language models' performance highlights areas where further improvements in model architecture and training strategies may be needed.

7. References

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134.

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.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2730–2741.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

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.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan S She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2022. Event knowledge in large language models: the gap between the impossible and the unlikely. *arXiv preprint arXiv:2212.01488*.

Soo-Hwan Lee and Shaonan Wang. 2023. Do language models know how to be polite? *Proceedings of the Society for Computation in Linguistics*, 6(1):375–378.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212.

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.

Stefan Palan and Christian Schitter. 2018. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Alicia Parrish, Amilleah Rodriguez, and Liina Pylkkänen. 2022. Non-local conceptual combination. *bioRxiv*.

Ellie Pavlick. 2022. Semantic structure in deep learning. *Annual Review of Linguistics*, 8:447–471.

Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2019. Towards sentence-level brain decoding with distributed representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7047–7054.

Shaonan Wang, Jiajun Zhang, Haiyan Wang, Nan Lin, and Chengqing Zong. 2020. Fine-grained neural decoding with distributed word representations. *Information Sciences*, 507:256–272.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221.

Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907.