Modeling Bilingual Sentence Processing: Evaluating RNN and Transformer Architectures for Cross-Language Structural Priming

Anonymous ACL submission

Abstract

This study evaluates the performance of Recurrent Neural Network (RNN) and Transformer models in replicating cross-language structural priming, a key indicator of abstract grammatical representations in human language processing. Focusing on Chinese-English priming, which involves two typologically distinct languages, we examine how these models handle the robust phenomenon of structural priming, where exposure to a particular sentence structure increases the likelihood of selecting a similar structure subsequently. Additionally, we use large language models (LLMs) to measure the crosslingual structural priming effect. Our findings indicate that transformers outperform RNNs in generating primed sentence structures, challenging the conventional belief that human sentence processing primarily involves recurrent and immediate processing, and suggesting a role for cue-based retrieval mechanisms. In general, this work contributes to our understanding of how computational models may reflect human cognitive processes in multilingual contexts.

1 Introduction

Existing studies show that Recurrent Neural Networks (RNN), particularly Gated Recurrent Unit models (GRU), have been pivotal in modeling human sentence processing (Frank et al., 2019). These models can explain phenomena like garden-path effects and structural priming. A garden-path effect occurs when a reader is led to interpret a sentence in a way that turns out to be incorrect, requiring reanalysis to understand the correct structure. Structural priming refers to the phenomenon where encountering a specific syntactic structure boosts the probability of generating or understanding sentences with a comparable structure (Pickering and Ferreira, 2008).

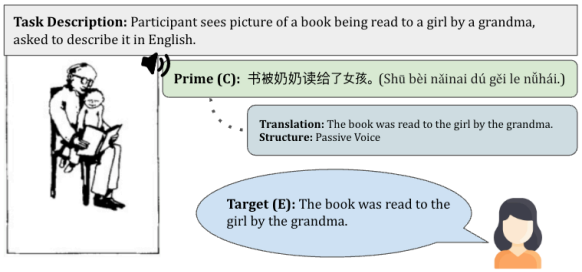

Figure 1: Cross-language structure priming of human participant: *C* denotes Chinese, *E* denotes English.

1.1 Cross-Linguistic Structural Priming

Prior experiments induce cross-linguistic structural priming by instructing bilingual participants to use two languages: presenting primes in one language and eliciting targets in another. These studies show that specific sentence structures in one language influences the use of similar structures in the other language (Hartsuiker et al., 2004).

Consider a case where a human participant reads a passive Chinese (C) sentence and is then asked to describe a separate picture in English (E) (see Figure 1). Here, the passive sentence C influences the structure of the target sentence E, leading the participant to use passive voice in their description.

Computational modeling studies have shown that RNNs exhibit structural priming effects akin to those observed in human bilinguals (Frank, 2021). These models process sequential information through recurrence, a feature thought to resemble human cognitive processing. The emergence of such priming effects in language models suggests that they develop implicit syntactic representations that resemble those employed by human language systems (Linzen and Baroni, 2021).

However, the transformer model, which uses self-attention mechanisms instead of recurrence, challenges this notion. The transformer's ability to directly access past input information, regardless of temporal distance, offers a fundamentally different approach from RNNs. The effectiveness

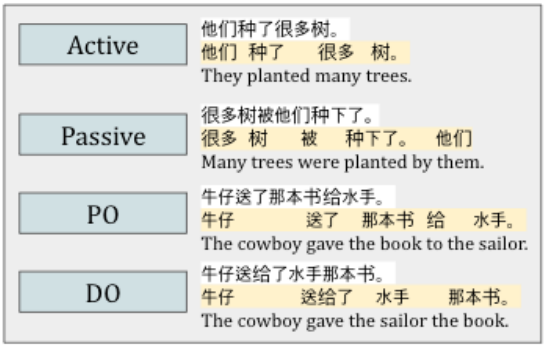

Figure 2: Example of Active, Passive, Propositional Object (PO), and Double Object (DO). White highlighted sentence is original Chinese sentence, and yellow highlighted Sentence is word-to-word mapping between Chinese and English.

of transformers and recent large language models (LLMs) in various NLP tasks makes us wonder if they can emulate RNNs in modeling cross-language structural priming.

1.2 Prior Studies

The current study is inspired by two prior studies. Merkx and Frank (2021) compare transformer and RNN models' ability to account for measures of monolingual (English) human reading effort. They show that transformers outperform RNNs in explaining self-paced reading times and neural activity during reading English sentences, challenging the widely held idea that human sentence processing relies on recurrent and immediate processing. However, the study is monolingual and English-centric. Frank (2021) investigates cross-language structural priming, finding that RNNs trained on English-Dutch sentences account for garden-path effects and are sensitive to structural priming, within and between languages.

1.3 The Current Study

Our study builds upon these two studies, comparing RNNs and transformers for their ability to model cross-language structural priming. We use a different metric for structural priming. Frank (2021) trains models on comprehension, where a longer response time indicates greater difficulty in understanding the new sentence, indicating a weaker priming effect. In contrast, our models are trained for production—the structure of the generated sentences is compared with that of the input sentence to assess the presence of a priming effect.

There are Chinese equivalents to passive *Many trees were planted by them.* and active *They planted many trees.*, as well as prepositional objects *The cowboy gave the book to the sailor.* and double objects *The cowboy gave the sailor the book.* as shown in Figure 2. In our study, the input sentence is in Chinese and system output is an English version of the sentence. BLEU scores are calculated between the system output English sentence and the English sentence that share structure with the the Chinese input—the "correct answer" as well as an "incorrect" answer. We then calculate the difference between two BLEU scores, as depicted in Figure 3.

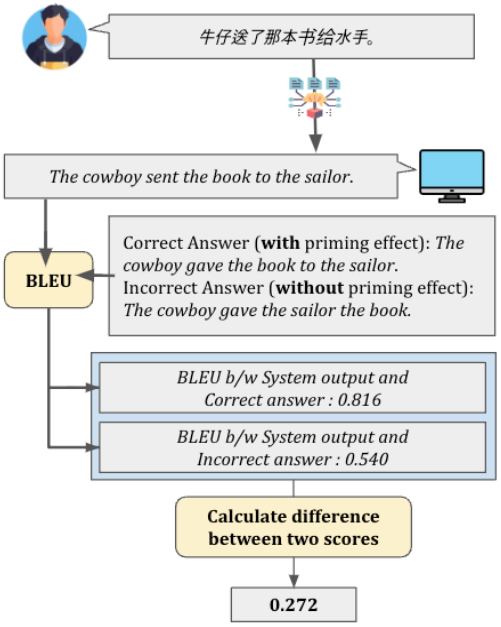

Figure 3: Example of test phase and evaluation process.

Another novel aspect of our study is that the two chosen languages are from distinct language families, challenging the models to develop abstract representations for structurally different forms.

2 Data Preparation

We select and process a Chinese-English corpus which contains 5.2 million Chinese-English parallel sentence pairs (Xu, 2019).[1]

We employ a DataLoader [2] to facilitate batch processing, transforming text into token IDs suitable for model interpretation. We then use the Helsinki-NLP tokenizer (Tiedemann and Thottingal, 2020)[3]

[1]The source can be found at https://drive.google.com/file/d/1EX8eE5YWBxCaohBO8Fh4e2j3b9C2bTVQ/view?pli=1

[2]Our Dataloader is supported by PyTorch, referencing its license located at https://github.com/pytorch/pytorch/blob/main/LICENSE

[3]Helsinki-NLP is licensed under the MIT license. For more details, see here: https://github.com/Helsinki-NLP/Opus-MT/blob/master/LICENSE

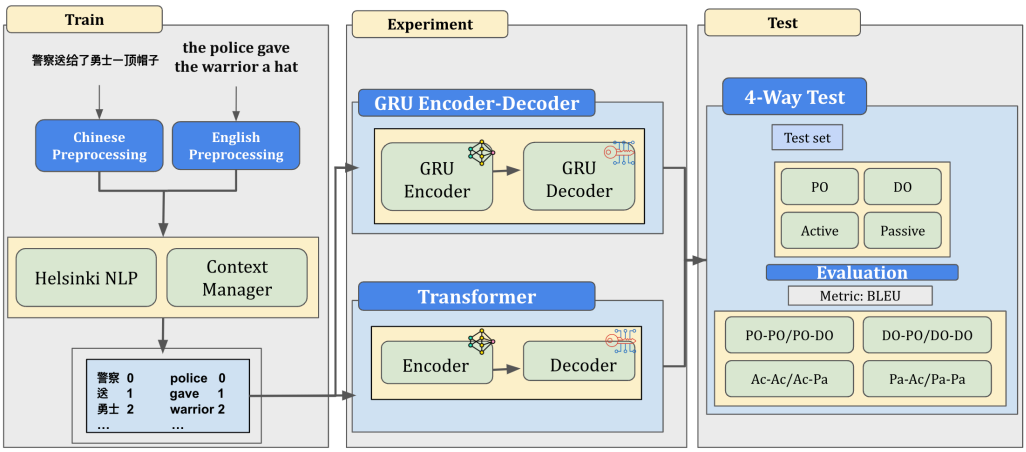

Figure 4: The overarching workflow of the study is illustrated as follows. PO refers to Propositional Object, DO refers to Double Object, Ac refers to Active and Pa refers to Passive. During the training phase, we preprocess the raw bilingual data through several steps to generate token pairs. In the experiment phase, we employ transformer and RNN-based encoder-decoder architectures. In the testing phase, we evaluate the model's performance using four different sentence structures and assess the output with the BLEU metric.

to map Chinese to English, accommodating over a thousand models for diverse language pairs.

The tokenizer, by default, processes text according to source language settings. To encode target language text, the context manager as a target tokenizer must be used. Without this, the source language tokenizer would be applied incorrectly to the target text, leading to poor tokenization results, such as improperly splitting words unrecognized in the source language.

In sequence-to-sequence models, setting padding tokens to -100 ensures they are ignored during loss calculations. This setup is crucial for effective model training, allowing for precise adjustment of model parameters based on the tokenized input and target sequences. Properly formatting the data through this preprocessing step facilitates optimal training outcomes.

We also design a test dataset. Initially, 5 sentences for each of the 4 types of sentence structures (Active Voice, Passive Voice, Prepositional Object, and Double Object) are sampled from Cross-language Structural Priming Corpus (Michaelov et al., 2023). Then, we employ a LLM, ChatGPT 3.5 (OpenAI, 2024), to augment the data. By providing the following prompt as one shot learning, we expand each set to 30 sentences, resulting in a total of 120 sentences for our test dataset:

> Generate 30 sentences with the following structure: *The cowboy gave the book to the sailor.* Replace all the words while keeping the sentence structure the same.

Correspondingly, in our test set, each Chinese sentence is paired with a correct and an incorrect English sentence.

3 Language Models

We implement both a transformer model and an RNN model to handle sequence-to-sequence tasks using the encoder-decoder architecture. (See Experiment of Figure 4) This architecture supports the processing of both input sequences and output sequences of varying lengths, which is crucial for accommodating sentences with different structures yet similar meanings. This section explores why these language models can assist us identify structural priming. We train and test our RNN model and transformer using AMD EPYC 75F3 8-Core Processor and 1 NVIDIA A100 GPU.

3.1 Multi-head Attention in Transformer

In the transformer model, we use the self-attention mechanism (AttModel) to capture sentence structure. This mechanism identifies dependencies between different positions and adjusts the representation of each word based on its relationship with others, thus facilitating the learning of sentence structure. Following Vaswani et al. (2017),

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where Q, K, V are obtained through linear transformations of an input sequence of text, each with its own learnable weight matrix. In the encoder part

of model, Q, K, V comes from the same source sequence, while in the decoder part, Q comes from the target sequence, and K and V come from the output sequence of the encoder. Since the computation of Q, K, and V requires processing the entire input sentence, the model can simultaneously focus on all positions and capture the structure of the sentence.

In the decoder part of the transformer model, the use of multiple attention heads allows for the capture of diverse levels of sentence features, leading to a more comprehensive representation of sentence structure. Each attention head specializes in capturing specific semantic relationships, such as word dependencies and distance relationships.

This approach enhances the model's ability to comprehend the intricacies of sentence structure. The equation is as follows:

$$\mathrm{MH}(Q, K, V) = \mathrm{Concat}(head_1, \ldots, head_h) \cdot W^O \quad (2)$$

where W^O is the weight matrix we need to train, and $head_1, \ldots, head_h$, computed through equation 1, represent the attention weights of each head (we choose to use 8 heads). Concat is the operation of joining tensors along their last dimension.

We also focus on selecting the positional encoding method. While the common method involves using sine and cosine functions, we opt for learnable positional embedding because we believe this approach offers more advantages for learning structural priming because it helps our model better understand and encode the relative positions of words within a sentence.

In contrast to the fixed positional encoding, learnable positional embeddings assign different weights to different positions, emphasizing the relevant positional information that contributes to the priming effect. This enables the model to capture more intricate positional relationships and dependencies specific to the task of structural priming.

3.2 GRU Encoder and GRU Decoder

Some studies (Zhou et al., 2018) show that RNNs can preserve sentence structure and facilitate identification of structural priming environment. Their sequential nature allows them to process input tokens based on a contextual understanding of the entire sentence. As each token is processed, the RNN's hidden state is updated, retaining information about preceding tokens and their contextual relevance. This sequential processing enables the model to capture word dependency relationships, thereby preserving the structural integrity of the sentence. Summarizing:

$$\mathrm{State}(dh_i, c_i), p = f(\mathrm{State}(dh_{i-1}, c_{i-1}), m) \quad (3)$$

where function f refers to the hidden layer of the RNN model, which is a neural network. It takes the previous layer's State i-1 and the output vector from the previous time step m as input, and outputs the next layer's State i and prediction value p until it encounters the termination symbol. In this state, dh signifies the hidden state of the RNN unit in decoder, tasked with capturing pertinent information gleaned from the input sequence. In the initial decoder step, dh embodies the final output state of the encoder. In subsequent decoder steps, dh denotes the preceding RNN unit's output.

To address the challenge of not being able to retain the entire sentence structure, we introduce the attention mechanism. This feature of the RNN model enables it to focus more on the parts of the input sequence that are most relevant to the current output, thereby enhancing prediction accuracy. Its potential for predicting structural patterns stems from its capability to capture dependencies within sequential data and to exploit these dependencies for prediction. As shown in equation 3, c denotes the attention. The calculation of c is as follows:

$$\alpha_i = g(eh_i, dh_0) \quad (4)$$

As before, dh_0 denotes the final state of the encoder and eh signifies the hidden state of the each RNN unit in encoder. Function g is used to calculate the weight $alpha_i$ of eh_i in the final state dh_0. As a result, we obtain the attention c by combining all previous states:

$$c_i = \sum(\alpha_i * dh_i) \quad (5)$$

calculated by summing the products of the weight α and the state in decoder dh.

Our study utilizes a variant of RNNs: the Gated Recurrent Unit (GRU). The GRU encoder and GRU decoder incorporate gating mechanisms, which can effectively manage long-distance dependencies and avoid the vanishing gradient problem. Additionally, GRUs possess fewer parameters and demonstrate higher computational efficiency.

Following Dey and Salem (2017), we define the gate mechanism in two parts:

- Update Gate:
  $z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$

The update gate z_t in the encoder controls the blending of the current input x_t and the previous hidden state h_{t-1}. The update gate z_t in the decoder regulates the interaction between the current input and the previous decoder state. This allows the model to selectively incorporate relevant information from the input when generating the output.

- Reset Gate:
  $r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$

The reset gate r_t in the encoder regulates the interaction between the current input x_t and the previous hidden state h_{t-1}. The reset gate r_t in the decoder governs how the current input interacts with the previous decoder state. This allows the model to selectively forget certain parts of the input information captured by the encoder, enabling the decoder to generate outputs that are less influenced by outdated information from the input sequence.

4 Experimental Setup

To assess the effectiveness of our model in Chinese-English, we adopt the standard bilingual evaluation understudy (BLEU) metric (Papineni et al., 2002), which ranges from 0 to 1, indicating the similarity of predicted text against target text:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Here, N is the maximum n-gram order (typically 4), w_n is the weight assigned to each n-gram precision score, ($\sum_{n=1}^{N} w_n = 1$) p_n is the precision score for n-grams of order n, and BP is the brevity penalty which penalizes shorter results.

After generating predicted outcomes and assembling a test set, we analyze the relationship between predictions and four types of reference sentences: (1) correct mappings with the same structure; (2) semantically similar but structurally different sentences; (3) semantically different but structurally identical sentences; and (4) sentences that differ both semantically and structurally.

We categorize the comparisons into two distinct groups based on semantic similarity. In the first category, encompassing sentences with identical meanings, we hypothesize that effective structural priming would result in higher BLEU scores between the predicted sentences and the reference sentences, with the same structure compared to those with different structures. This comparison

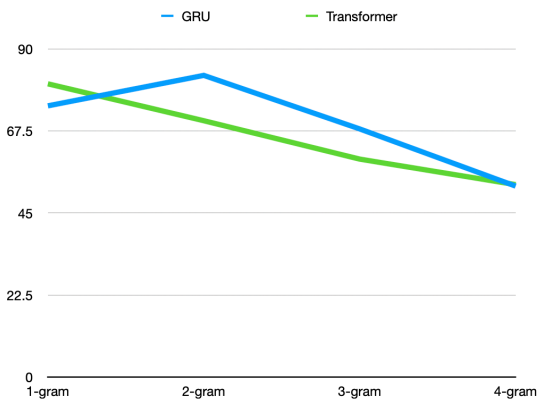

Figure 5: BLEU Score for standard structural priming. Comparison of ground truth datasets for testing and calibration.

aims to establish whether the model exhibits a preference for reproducing structures that are syntactically aligned with the ground truths when the semantic content is constant.

The second category, which involves sentences that differ in meaning, is particularly crucial for demonstrating structural priming, as it eliminates the influence of semantic similarity. If sentences with identical structures receive higher BLEU scores compared to those with different structures, it would strongly suggest that the model's predictions are influenced by the structural aspects of the input, regardless of semantic changes.

Through this methodology, we seek to rigorously test for structural priming outputs, offering insights into how the models process and replicate structural properties of language.

5 Results and Analyses

We present the performance of the GRU-based RNN and standard transformer model (Vaswani et al., 2017) and then demonstrate their crosslingual structural priming effect in Chinese-English bilingual scenarios. We also present our insights regarding the performance of open-source large language models on the same dataset accordingly.

5.1 Structural Priming Performance

Our comparative analysis reveals that, although both models achieve competitive BLEU scores, the transformer model shows a slight edge in handling complex sentence structures. Figure 5 shows that, when the training dataset is sufficiently large, both models attain high predicted BLEU scores for standard structured sentence segments.

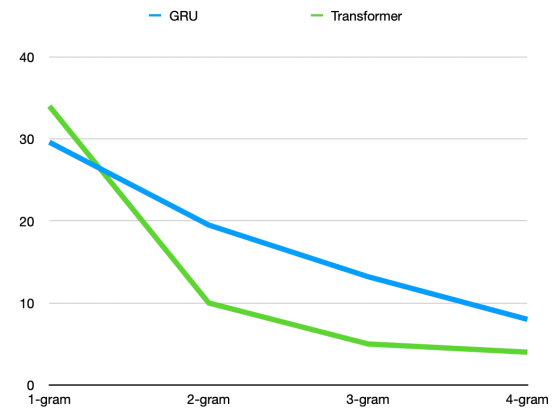

Figure 6: BLEU Score for wrong priming. Comparison between predictions for cross-language priming via average BLEU Score.

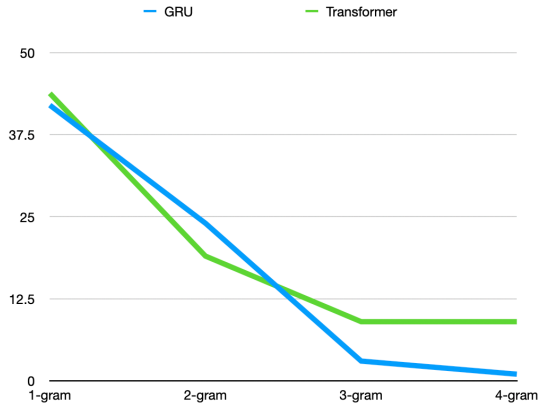

Figure 7: BLEU Score for correct priming. Comparison between predictions for opposite cross-language priming via average BLEU Score.

5.2 Crosslingual Structural Priming Effect

Through our examination of crosslingual structural priming, we observe a noteworthy pattern: both models facilitate the use of target-language syntactic structures influenced by the source language. However, the transformer model displays a more pronounced priming effect, indicating a potential edge in mimicking human-like syntactic adaptation in bilingual contexts.

Figure 6 and Figure 7 show the BLEU scores for machine-generated predictions with correct or opposite priming test sets. From these we gain insights into model performance. Specifically, we evaluate the similarity levels between the model predictions and the correct priming test sets (e.g., Active-Active, DO-DO) as well as the opposite priming test sets (e.g., Active-Passive, PO-DO). Higher BLEU scores against the correct priming test sets indicate that the model predictions align more closely with the appropriate structural priming, whereas higher scores against the opposite priming test sets suggest deviations from the expected priming behavior.

The results reveal that when evaluated against the correct priming test sets, the transformer model exhibits similar levels to GRU (see Figure 6), with slight improvements observed as the n-gram size increases. Conversely, in comparison to opposite priming, GRU generally outperforms the transformer (see Figure 7). Given that this comparison involves what is termed as "incorrect" priming, GRU aligns more closely with the opposite priming test set. Since transformer shows a larger gap between BLEU score (correct) and BLEU score (wrong), We infer that the transformer adheres more closely to the appropriate structural priming.

In a previous study, Michaelov et al. (2023) examine the presence of structural priming by comparing the proportion of target sentences produced after different types of priming statements. Similarly, for each experimental item in our study, we prime the language model with a specific sentence and calculate the normalized probabilities for the two target sentences. These normalized probabilities are computed as follows:

First, calculate the raw probability of each target sentence given the priming sentence:

$$P(\text{DO Target}|\text{DO Prime})$$
$$P(\text{PO Target}|\text{PO Prime})$$
$$P(\text{DO Target}|\text{PO Prime})$$
$$P(\text{PO Target}|\text{DO Prime})$$

And the same method for:

$$P(\text{Active Target}|\text{Active Prime})$$
$$P(\text{Passive Target}|\text{Passive Prime})$$
$$P(\text{Active Target}|\text{Passive Prime})$$
$$P(\text{Passive Target}|\text{Active Prime})$$

These probabilities are then normalized to calculate the conditional probability of the target sentence if the model output is one of the two target sentences. Taking DO | PO as example:

$$P_N(\text{Target}|\text{Prime}) = \frac{P(\text{Target}|\text{Prime})}{P(\text{DO Target}|\text{Prime}) + P(\text{PO Target}|\text{Prime})}$$

Since the sum of the normalized probabilities of the two target sentences is 1, we only need to consider the probability of one target type and compare between different priming types. The reason is that the probability of another target type can be derived from this, i.e. $P_N(\overline{\text{Target}}|\text{Prime}) = 1 - P_N(\text{Target}|\text{Prime})$. By considering only one

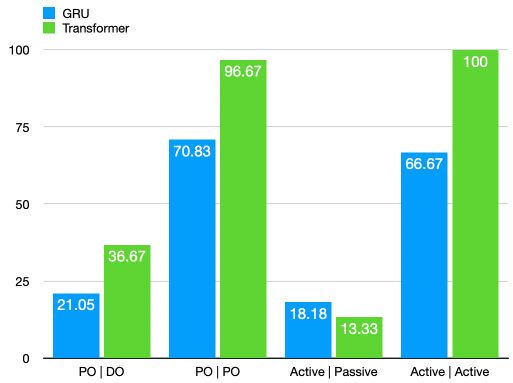

Figure 8: Priming Effect per Chunk: Proportion of correct cross-language priming chunks in the machine prediction results.

goal type, we can directly compare the priming effects of the two priming types on that goal type, which is the main focus of analysis in structural priming research. The quantitative comparative findings depicted in Figure 8 derived from the sentence chunk dimension reveal that the transformer model generally outperforms GRU. Through a horizontal examination of priming structural types, it is evident that machine predictions exhibit superior performance with respect to active/passive structures compared to those of PO/DO.

The trained transformer model is only exposed to the Chinese-English dataset. Prior research (Michaelov et al., 2023) has shown that LLMs can mimic human language structural priming effects in various scenarios, both in within-language and in crosslingual experiments. However, there is a lack of evidence regarding the effectiveness of such multilingual language models in demonstrating Chinese-English structural priming effects. To address this, we adopt XGLM model proposed by Lin et al. (2022) [4] and evaluate their performance using the normalized score defined above, on the same set of tasks designed for RNN and transformers. Among the four categories in our study, we find this language model family exhibits greater sensitivity in demonstrating structural priming effects in passive and prepositional tasks (see Figure 9), with the effect being more noticeable in the former case.

[4]XGLM is developed by Facebook Research and is available under the MIT license. For more details, see the license at https://github.com/facebookresearch/fairseq/blob/main/examples/xglm/README.md

6 Discussion

This study evaluates cross-language structural priming effects in RNN and transformer models in the context of Chinese-English. We find evidence for abstract crosslingual grammatical representations in these models, which operate similarly to those found in prior research.

6.1 Conclusions

Our results show that BLEU scores decrease as the length of n-grams increases, a trend that is consistent with existing findings in sentence-similarity evaluation (He et al., 2022). Longer n-grams such as bigrams and trigrams, capture more specific linguistic contexts, making exact matches less likely unless the target sentence is very precise. Moreover, any minor errors in word choice or sequence can disrupt the alignment of these longer n-grams.

Importantly, our results indicate that transformer models outperform RNNs in modeling Chinese-English structural priming, a finding that is intriguing given prior research. Traditionally, RNNs have been effective in modeling human sentence processing, capable of explaining phenomena such as garden-path effects and structural priming through their sequential processing capabilities, which are thought to mirror aspects of human cognitive processing (Frank, 2021).

This superiority of transformers raises questions about the efficacy of RNNs as human sentence processing models, especially if they are surpassed by a model considered less cognitively plausible. However, it is possible to interpret the results as supportive of the cognitive plausibility of transformers, particularly due the attention mechanism.

While the concept of unlimited working memory in transformers is viewed as implausible, some researchers argue that actual human working memory capacity is much smaller than traditionally estimated—limited to only two or three items. They suggest that language processing involves rapid, direct-access retrieval of items from memory (Lewis et al., 2006), a process compatible with the attention mechanism in transformers. This mechanism assigns weights to previous inputs based on their relevance to the current input, which aligns with cue-based retrieval theories, indicating that memory retrieval is influenced by the similarity of current cues to stored information (Parker and Shvartsman, 2018).

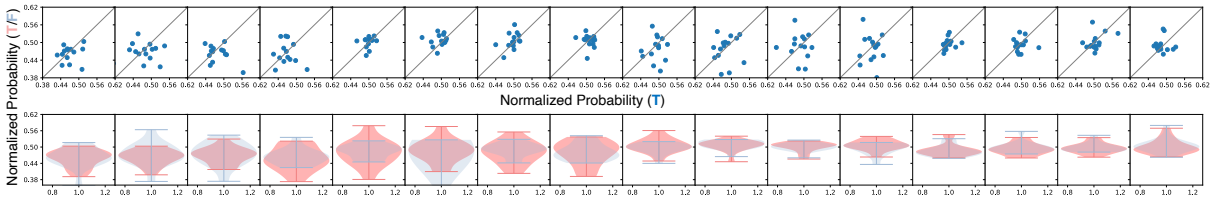

Figure 9: Crosslingual structural priming effect for large language model, XGLM (Lin et al., 2021), in the context of Chinese-English with various sentence types. Top: Normalized probability of true target verus false target prediction based on the next word logit value. Bottom: summary of each comparison shown in the same raw above. From left to right: XGLM with parameters of 564M, 1.7B, 2.9B, 4.5B on active, passive, prepositional, and Double tasks.

6.2 Future Directions

A promising future direction involves developing a model capable of generating sentences based on new semantic concepts and thematic roles before and after priming. Although this endeavor presents challenges, it holds the potential to mitigate the lexical boost effect (see Limitations).

Shifting our focus from production to comprehension could also be fruitful. By measuring the surprisal levels in models, we can gain insights into how structural priming influences model comprehension, as suggested in recent studies (Merkx and Frank, 2021). Surprisal, in information theory and psycholinguistics, quantifies the unexpectedness of a word in a given linguistic context. Lower suprisal values indicate greater probability, i.e., consistently lower surprisal levels at structurally complex points in sentences that follow a priming example would suggest effective preparation by the priming process. This method offers a way to explore how structural priming impacts language processing in models, without the confounding effects of repeated vocabulary.

Additionally, there is evidence suggesting an inverse relationship between the frequency of linguistic constructions and the magnitude of priming effects observed with those constructions (Jaeger and Snider, 2013; Kaschak et al., 2011). For example, the double object (DO) construction is more common in American English than the prepositional object (PO) construction (Bock and Griffin, 2000). Studies have shown that the less frequent PO construction exhibits stronger priming effects compared to the more frequent DO construction (Kaschak et al., 2011). This aligns with theories of implicit learning in structural priming, where more frequently encountered structures are less "surprising" to the language system and thus generate weaker priming effects.

To delve further into this, training models on corpora consisting of American versus British English, which differ in their construction frequencies, could reveal whether a similar inverse frequency effect is observed in computational models. This approach would help illuminate the dependency of structural priming on construction frequency, potentially providing deeper insights into how implicit learning processes are modeled computationally.

Limitations

A limitation of the current study is that Chinese-English priming effect of the models is not compared with human data. We equate the models' ability to replicate cross-language priming with the structural "correctness" of their outputs, yet empirical studies indicate that even humans do not achieve full priming rate (Hsieh, 2017). Therefore, it is conceivable that if the models' outputs are compared directly to human data, RNNs might more closely resemble human performance. This limitation highlights an area for future research, which could involve direct comparisons to human priming data to better assess the models' fidelity to human language processing.

A further limitation is that our models are not capable of generating sentences based on novel word concepts and thematic roles, such as the picture naming task in Figure 1. Consequently, some critics may argue that what our models essentially do is translate from Chinese to English without generating new semantic content, as the semantic information remains consistent from the priming sentence to the output sentence.

Despite these critiques, we maintain that the current study design still validly assesses the priming effect. This is because the models must choose which sentence structure to use from among vari-

ous structures that share the same semantic content, a choice influenced by the priming effect.

Nevertheless, we acknowledge that our design is susceptible to the "lexical boost" effect, where the structural priming effect is intensified when the same lexical head is repeated in both the prime and target sentences (Pickering and Branigan, 1998). For instance, if the target sentence is "Alice gave Bob a book," the priming effect is more pronounced if the prime sentence was "Carl gave Danis a letter" rather than "Alice showed Bob a book." Given that the semantic content remains constant across the prime and output sentences in our study, the observed priming effect is artificially strengthened compared to what might be observed in a pure priming task.

Another aspect worth discussing is the significance of using LLMs to simulate human language processing efforts. As highlighted in the introduction, the ultimate goal is to deepen our understanding of how the human brain functions, assuming that models which appear more human-like externally might also mirror human cognitive processes internally. However, one might question the validity of using LLMs for this purpose. Given that these models often function as "black boxes," their internal operations remain largely opaque. Despite their impressive computational abilities, the lack of transparency means that even if they outperform more interpretable models, they do not necessarily enhance our understanding of brain function.

Previous studies argue that crosslingual structural priming might be affected by the asymmetry of training sources in certain language pairs (Michaelov et al., 2023). By measuring the probability shifts for source and target sentences, we find such multilingual auto-regressive transformer language models display evidence of abstract structural priming effect, although their performance varies across different scenarios.

Ethical Statement

The current study adheres to the ethical standards set forth in the ACL Code of Ethics. The training dataset used in this research is open, publicly available, and does not include demographic or identity characteristics (Xu, 2019).

Potential risks may arise from the fact that translations in the training data (a Chinese-English parallel sentence pair dataset) may not always be perfectly equivalent. Some words may carry cultural nuances that differ between Chinese and English. For example, the terms "和尚" (heshang) and "尼姑" (nígū), translated as "monk" and "nun," have specific cultural connotations in Chinese that differ from the perception of a "monk" in Western contexts, which is typically associated with Christian monasticism. These roles in Chinese Buddhism embody cultural and social aspects not fully captured by the Western terms, potentially leading to a loss of cultural meaning in translation.

Furthermore, while ChatGPT has been used to expand the test dataset, the authors have manually verified the output to ensure it remains unbiased.

The potential risk of misuse of the computational model is low, as the encoders and decoders are designed to perform straightforward translation tasks and do not have the capability to self-generate harmful content.

References

Kathryn Bock and Zenzi M. Griffin. 2000. The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General*, 129(2):177–192.

Rahul Dey and Fathi M Salem. 2017. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE.

Stefan Frank. 2021. Cross-language structural priming in recurrent neural network language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Stefan L Frank, Padraic Monaghan, and Chara Tsoukala. 2019. Neural network models of language acquisition and processing. In *Human language: From genes and brain to behavior*, pages 277–293. MIT Press.

Robert J. Hartsuiker, Martin J. Pickering, and Eline Veltkamp. 2004. Is Syntax Separate or Shared Between Languages?: Cross-Linguistic Syntactic Priming in Spanish-English Bilinguals. *Psychological Science*, 15(6):409–414.

Jia-Wei He, Wen-Jun Jiang, Guo-Bang Chen, Yu-Quan Le, and Xiao-Fei Ding. 2022. Enhancing N-Gram Based Metrics with Semantics for Better Evaluation of Abstractive Text Summarization. *Journal of Computer Science and Technology*, 37(5):1118–1133.

Yufen Hsieh. 2017. Structural priming during sentence comprehension in Chinese–English bilinguals. *Applied Psycholinguistics*, 38(3):657–678.

T. Florian Jaeger and Neal E. Snider. 2013. Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, 127(1):57–83.

Michael P. Kaschak, Timothy J. Kutta, and John L. Jones. 2011. Structural priming as implicit learning: Cumulative priming effects and individual differences. *Psychonomic Bulletin & Review*, 18(6):1133–1139.

Richard L. Lewis, Shravan Vasishth, and Julie A. Van Dyke. 2006. Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, 10(10):447–454.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. 2021. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O' Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot Learning with Multilingual Generative Language Models. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

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

Danny Merkx and Stefan L. Frank. 2021. Human Sentence Processing: Recurrence or Attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22, Online. Association for Computational Linguistics.

James A. Michaelov, Catherine Arnett, Tyler A. Chang, and Benjamin K. Bergen. 2023. Structural Priming Demonstrates Abstract Grammatical Representations in Multilingual Language Models. *arXiv preprint arXiv:2311.09194*. Publisher: [object Object] Version Number: 1.

OpenAI. 2024. Gpt-3.5 turbo documentation. Accessed: 2024-06-10.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Dan Parker and Michael Shvartsman. 2018. The cue-based retrieval theory. *Language Processing and Disorders*, page 121.

Martin J. Pickering and Holly P. Branigan. 1998. The Representation of Verbs: Evidence from Syntactic Priming in Language Production. *Journal of Memory and Language*, 39(4):633–651.

Martin J Pickering and Victor S Ferreira. 2008. Structural priming: a critical review. *Psychological bulletin*, 134(3):427.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT —Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Bright Xu. 2019. Nlp chinese corpus: Large scale chinese corpus for nlp.

Yi Zhou, Junying Zhou, Lu Liu, Jiangtao Feng, Haoyuan Peng, and Xiaoqing Zheng. 2018. Rnn-based sequence-preserved attention for dependency parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.