※試験場への持込
１．不可
２．可　イ．辞書
ロ．その他
（英和・和英辞典各１冊）

２０２６年度【Ｉ期】　玉川大学大学院脳科学研究科脳科学専攻
博士課程後期入学試験問題

科目名	外国語（英語）	受験番号		氏名	

次の文章は，論文　**Kumar, Sreejan et al. 2024. "Shared Functional Specialization in Transformer-Based Language Models and the Human Brain." Nature Communications 15 (1): 5523.**からの引用である．次の文章を読んで，以下の問いに答えよ．

Traditionally, neuroimaging research has used targeted experimental manipulations to isolate particular linguistic computations—for example, by manipulating the presence/absence or complexity of a given syntactic structure—and mapped these computations onto brain activity in controlled settings. (a)While these findings laid the groundwork for a neurobiology of language, they have limited generalizability outside the laboratory setting, and it has proven difficult to synthesize them into a holistic model that can cope with the full complexity of natural language. This has prompted the field to move toward more naturalistic comprehension paradigms. However, these paradigms introduce their own challenges: principally, how to explicitly quantify the linguistic content and computations supporting the richness and expressivity of natural language.

In recent years, the field of natural language processing (NLP) has been revolutionized by a new generation of deep neural net- works capitalizing on the Transformer architecture. (b)Transformers are deep neural networks that forgo recurrent connections in favor of layered "attention head" circuits, facilitating self- supervised training on massive real-world text corpora. Following pioneering work on word embeddings, the Transformer architecture represents the meaning of words as numerical vectors in a high-dimensional "embedding" space where closely related words are located nearer to each other. However, while the previous generation of embeddings assign each word a single static (i.e., non-contextual) meaning, Transformers process long sequences of words simultaneously to assign each word a context-sensitive meaning. The core circuit motif of the Transformer—the attention head—incorporates a weighted sum of information exposed by other words, where the relative weighting "attends" more strongly to some words than others. The initial embeddings used as input to the Transformer are non-contextual. Within the Transformer, attention heads in each layer operate in parallel to update the contextual embedding, resulting in surprisingly sophisticated representations of linguistic structure.

The success of Transformers has inspired a growing body of neuroscientific work using them to model human brain activity during natural language comprehension. (c)These efforts have focused exclusively on the "embeddings"—the Transformer's representation of linguistic content—and have largely overlooked the "transformations"—the actual computations performed by the attention heads. Although no functionally-specific language modules are built into the architecture at initialization, recent work in NLP has revealed emergent functional specialization in the network after training55,56.

That is, particular attention heads are shown to selectively implement interpretable linguistic operations. For example, attention head 10 in the eighth layer of BERT appears to be specialized for resolving the direct object of a verb (e.g., in "the boy in the yellow coat greeted his teacher", the verb "greeted" attends to "boy"), whereas head in the same layer closely tracks nominal modifiers (e.g., attending to "coat" in the phrase modifying "boy"). This is in contrast to probabilistic syntactic parsers, which learn to reproduce a predefined set of syntactic labels to construct parse trees. The trans- formations do not explicitly disentangle syntax from the meaning of words and do not rely on predefined labels; instead they learn to approximate whatever contextual structures are useful for accurately predicting words in real-world text. Although the individual heads that implement these computations operate independently, in parallel, their transformations are ultimately "fused" together to form the resulting embedding. Thus, unlike the embeddings, the transformations at a given layer can be disassembled into the specialized computations performed by the constituent heads. These transformations are of particular theoretical interest, because they are the unique component of the circuit that allows information to flow between words: whatever syntactic or contextual information impacts the meaning of the current word is introduced solely via the transformations.

In the current work, we argue that the headwise transformations—the functionally specialized contextual computations implemented by individual attention heads—can provide a complementary window onto linguistic processing in the brain. A neurocomputational theory of natural language processing must ultimately specify how meaning is constructed across words. (d)The Transformer architecture provides explicit access to a candidate mechanism for quantifying how the meaning of past words is incorporated into the meaning of the current word. If this is an important part of human language processing, these transformations should provide a good basis for modeling human brain activity during natural language comprehension. We extract transformations from the widely-studied BERT model and use encoding models to evaluate these trans- formations against several other families of linguistic features in terms of predicting brain activity during natural language comprehension. We find that the transformations perform comparably to the embeddings, and generally outperform both non-contextual embeddings and classical syntactic annotations—suggesting that the contextual information extracted from surrounding words is surprisingly rich. In fact, transformations at earlier layers of the model account for more unique variance in brain activity than the embeddings themselves. Finally, we disassemble these transformations into the functionally specialized computations performed by individual attention heads. We find that certain properties of the heads, such as look-back distance, dominate the mapping between headwise transformations and cortical language ears. We also find that, for some language regions, headwise transformations that preferentially encode certain linguistic dependencies also better predict brain activity.

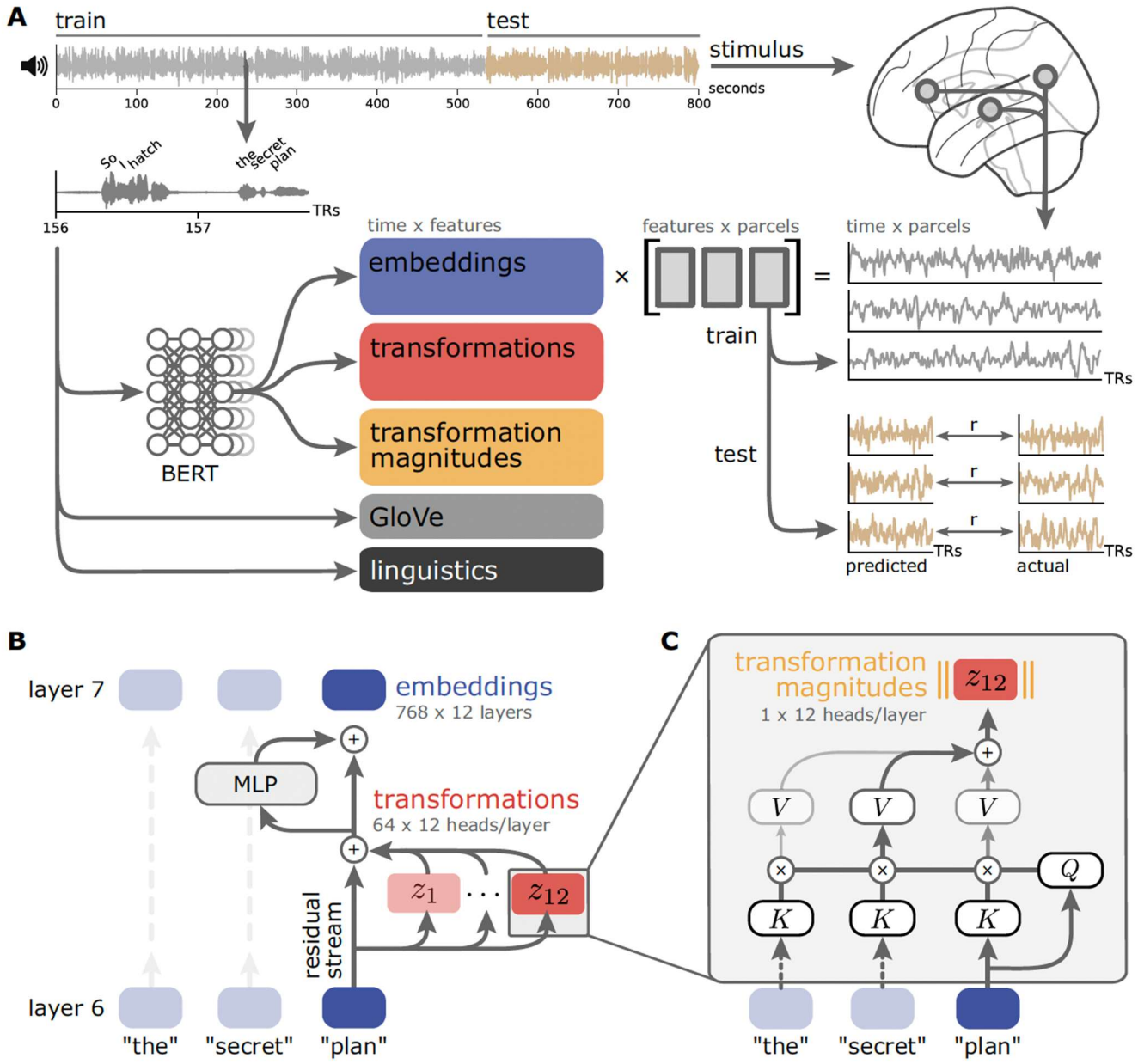

Fig.1: Encoding models for predicting brain activity from the internal com- ponents of language models.

補足情報

Transformer: 自己注意機構を積み重ねた最新の言語モデルの枠組み。RNN のような再帰結合を使わず、大量コーパスで学習する。

attention head: 文中の各語について「どの語にどれだけ注目するか」を計算し、その結果で表現ベクトルを更新する小さな担当ユニット

embeddings: 単語の意味を低次元ベクトル空間へ埋め込んだ表現．表現ベクトルとも呼ばれる．

transformations: Transformer の中で、各 **attention head** が「周りの語をどれだけ参照するか」を計算した結果として、いまの単語の表現ベクトルに加える文脈情報の更新のこと

context window: モデルが同時に参照できる範囲。長い窓ほど長距離の依存関係を扱える。

BERT: 文章の前後の文脈を読むことで、各単語の意味を数値ベクトルとして表す言語モデル

headwise transformations: ヘッド単位の変換；文章を読むとき、それぞれの **attention head** が、各単語の意味ベクトルに加える小さな更新のこと。

問 1

下線部**(a),(b),(c),(d)**をそれぞれ日本語に翻訳せよ

問 2

なぜ著者は **embeddings** だけでなく **transformations** を分析対象とすべきだと考えているのか？ 論文の記述に基づいて、日本語で根拠を説明せよ。

問 3

以下の**(1),(2)** のどちらか一つについて、あなた自身の考えを英語で **300** ワード程度で述べよ。
なお、答案用紙には選択した番号を明記せよ。

(1) **Transformer** における **attention heads** が生成する **transformations** と、人間の脳における言語処理は、どの点で共通し、どの点で本質的に異なるか？

(2) 今後の研究として、**Transformer** の仕組みと人間の脳をより深く比較するために、どのようなアプローチが必要か？