Do BERT's Embeddings Manifest Symmetry in Opposite Positional Statements?

Thomas W. Bush (ID:248201), Michael A. Awanah (ID:247099)

January 2024

1 Motivation and Introduction

The present study seeks to examine the ability of Large Language Models (LLMs) to represent statements about physical relationships between entities. We decide to focus on sentences containing opposite positional statements, such as: "The cat is to the left of the dog." and "The dog is to the right of the cat.". These two sentence represent the same spatial relationship, however, they describe it using a different point of reference.

Humans do not have the need to explicitly state both sentences for representing the mutual relationship between the entities, and they can easily change the point of reference when needed. This ability is particularly useful when they have to derive new spatial relationships. When another entity is added in relation to one of the previous, such as "The dog is to the left of the kid.", it is possible to use the mutual representation to derive the relationships between all the entities with more efficiency (two statements, instead of six).

In the following example we represent the symmetrical property of opposite positional statements as:

If X is left of Y, then Y is right of X.
If X is right of Y, then Y is left of X.

Derivation example:

Given: C is left of $D \Rightarrow D$ is right of C.
Given: D is left of $K \Rightarrow K$ is right of D.
Therefore, C is left of K and K is right of C.

C, D and K are the three entities of the previous sentences. Clearly, if humans would not use the symmetrical rule for positional statements, long spatial derivation would have to be explicitly stated.

Since 2017, when the transformer architecture has been applied for NLP tasks following (1), LLMs have become more competent in linguistics tasks, such as: respecting the hierarchy in a sentence (2), constructions (3) or abstractions (4).

These results, and the significant performances of these models on linguistics benchmarks (5), have made people think that LLMs were getting closer to human intelligence. In a recent review on the current abilities of LLMs it is argued that there is a difference between *linguistic competence* and *functional linguistics competence* (6). And if the former seems to be achieved by LLMs, the latter poses a difficult challenge for them. Functional linguistic competence entails more abilities than just producing correct sentences, such as: formal reasoning (7), social reasoning and common sense knowledge, all tasks where LLMs still struggle.

Spatial relationships are a particular case of formal reasoning. As we have seen above humans use the symmetry between opposite relationships to efficiently derive multiple relationships between entities without the necessity to explicitly represent them. This implicit derivation, which seems to rely on areas of the brain that fall outside the language's network (8), is not straightforward for AI models trained primarily on text data.

The object of this study is to test how LLMs represent spatial relationships, and whether their representations manifest the symmetrical property that

we have shown above. A better understanding in this domain would help to clarify the current limits of LLMs and the reason of why some of these tasks are tasks are still a challenge for them.

2 Literature

2.1 Main Works:

Symmetry in natural language is used to make systematic derivation about relations between entities (9). In verbs symmetry indicates a communal participation of the entities in the same action. This type of verbal symmetry is defined as: $x, y : R(x, y) \longleftrightarrow R(y, x)$ where x, y are the entities and R is the relation (9). In the case of symmetrical spatial relationships, it is not sufficient to permute the entities, but the relation must be inverted, $x, y : R(x, y) \longleftrightarrow R^{-1}(y, x)$.

Following we are going to review evidence on how humans and LLMs can represent this property.

2.1.1 How humans represent space:

We have seen that humans do not have the necessity to use language for expressing all the spatial relationship explicitly, because they are able to derive them intuitively.

One of the most important functions of language is to allow social cooperation efficiently (10), particularly it allows individuals to share perceptual maps to other individuals that did not have that perceptual experience. It follows that language communication would not be possible between individuals with different perceptual, cognition systems. Thus, spatial relations do not need to be explicitly stated because we rely on the recipient's ability to derive them. The human language network is extremely selective for language, and it does not participate in non-linguistic cognition (6). In particular spatial processing has been linked with a network in the hippocampal formation (11), that is active even during non-linguistic tasks.

Importantly, it has been proposed that the conceptual human system is profoundly shaped by physical properties (12); following this, it has been found that areas used for spatial cognition are also used for creating cognitive maps related to abstract concepts (13). It follows that during spatial communication language can be used as a proxy to elicit the recipient's spatial cognition area for making the correct derivations. In this way there is no need to state all the spatial relationships as they can be derived and improve efficient communication.

2.1.2 How LLMs represent space:

LLMs based on the transformer architecture use an encoder to represent the input sentence. The encoder is composed by a stack of identical layers, each of one containing two main layers: multi-head self attention mechanism and a position-wise fully connected feed-forward network. The key for the representation is the self attention mechanism (1), which enables the model to weight the importance of different words in a sentence. This architecture has allowed LLMs to be able to represent many linguistic features of a sentence that were associated with classic linguistic knowledge (14).

LLMs have shown promising evidence for being considered highly correlated with the human language network (15). They have also shown to be able to succeed in complex linguistics tasks that require a level of abstraction, for instance: dependencies between nonsensical terms (16), or respect hierarchical dependency between subject and verb, even if the verb is distant from the subject (17).

The notable results obtained by LLMs prompted researchers into exploring their abilities in non-linguistic domains (6). Relevant for our study, the spatial reasoning abilities of LLMs have been tested on multiple spatial task, such as: 2D path labelling, 3D trajectory labelling and spatial relationship identification (18). Other studies have identified challenges in LLMs for representing geometry and spatial relations (19).

Formal reasoning studies have shown how LLMs still struggle (20). It has been noticed how LLMs can improve their performances using a different prompting method, chain-of-though (21). However, the gap between linguistic and non-linguistic abilities of these models remain significant (6).

3 Research Question

This study aims to contribute to the literature about how LLMs represent spatial relations. We are especially interested in how symmetrical opposite positional statements are represented and whether LLMs include this property in their embeddings in a way that reflects how humans represent them.

Symmetry has been defined a key property of human language (9), we have seen that symmetry can make language communication more efficient and it is probably supported by a brain area outside the language network.

The expanding usage of LLMs outside the language domain has made it important to test how these models represent knowledge and if their representation is similar or not to what humans do.

For our test we have considered a famous encoder: BERT (22), a model that has achieved promising result in linguistic knowledge representation (14). We asked how BERT represents spatial properties that are symmetrical.

We have considered a simplified version of positional statements, such as: "A is at the left of B.", where A, B are two variable entities, and "left of" is the spatial relationship that links them.

These type of statements are able to convey explicit information about the state of "A" in relation to "B", this information can be represented by a vector, $V(S_1)$, however, there is also an implicit type of information about the state of "B" in relation to "A". This information is the inverse relation between the two entities, and it can be represented by the statement "B is to the right of A". It appears to us that there must be a symmetry between the two statements that ease the derivation of one from the other. The conceptual representation of "left" should be symmetrical to "right" in the conceptual space.

4 Proposal

For our experiment we focused on how BERT handles these type of statements in producing its embeddings, and whether they manifest the same symmetry that we find physically and conceptually in humans. Our hypothesis is that if BERT is able to account for the symmetry between opposite positional statements, then the representation of "right" and "left" should manifest this property, as it must be represented in its embeddings.

5 Project

5.1 Methods:

5.1.1 Conceptual assumptions

LLMs treat entities as high-dimensional vectors, where the dimensions of the vector reflect different properties of the word. It has been proposed that adjectives can be represented as matrix that modify the noun vector in the semantic space (23). In this case the adjective adds a constant value to the noun vector to represent their union.

Following this idea, we conceptualize spatial relationships as a transformation matrix that takes the vector entities as input and produces a final vector of their relationship in space. Specifically, since inverse spatial relationships should cancel themselves, we posit that the matrix should be square, so invertible, and their dot product must produce an identity matrix.

5.1.2 Experimental set-up

We have worked by using a custom dataset (1) composed by composed by entries of this type: the full sentence containing the spatial relationship, entity number one, entity number two, the type of spatial relationship, either "left of" or "right of". Subsequently, we have encoded the spatial relationship as a label (0 for 'left of', 1 for 'right of').

The first step for our experiment was to probe BERT in order to find the layers that handle the actual spatial relationship and reduce possible noise in the extracted vector. To do so, we have created a simple binary classifier using Logistic Regression from the 'insert library'; the classifier went through each BERT's layer and made a prediction on whether the statement was of 'left of' or 'right of' type. Then

Sentence	Entity_1	Entity_2	Spatial_rel
The clock is left of the painting.	The clock	The painting	left of
A tree is left of the building.	A tree	The building	left of
The lamp is left of the television.	The lamp	The television	left of
A cat is sitting left of the dog.	A cat	The dog	left of
The coffee cup is left of the computer.	The coffee cup	the computer	left of

Table 1: Example of dataset

we measured the differences in accuracy for each layer added to find where the spatial information is stored.

We couldn't get the transformation matrix directly from BERT, so we thought that the matrix function could come from the weights of a neural network that was trained to mimic BERT's embedding process for spatial language. Thus, we constructed a neural network to take the concatenation of the embeddings of the two entities from BERT and output a single vector representing their spatial relationship.

The model has been trained by using as a loss function the cosine similarity between its outputs and BERT's embeddings. After the training phase, we tested the network with new sentences, and we extracted the weights as transformation matrix M for the spatial relation "left of".

From M we have produced the inverse matrix M^{-1} for the inverse spatial relation. From our conceptual hypothesis we believe that if BERT is able to represent the symmetry between opposite positional statements, then the inverse matrix would be appropriate for the production of the statements of the "right of" type. Thus, by replacing the weights of the previous model with the new matrix we have produced new embeddings containing the symmetrical relation.

To asses the whether these embeddings are similar to the one produced by BERT on the full sentence, we have measured the cosine similarity between the two.

We would expect an high cosine similarity value (> 0.8) between the embeddings if BERT manifest the property of symmetry between opposite positional statements.

5.2 Results:

The neural network trained to reproduce the "left of" statements from the two entity's vectors reached an accuracy of 0.94, which implies that it correctly learned how to transform the entities into the final embedding of their spatial relationship.

When we tested the embeddings produced by the same model with the new matrix M^{-1} for the weights with the one produced by BERT on the full sentences, we have observed a consine similarity value of: 0.001. A value that indicates that the embeddings produced are not similar.

6 Discussion

The results obtained suggest that LLMs do not represent opposite spatial relations symmetrically. This, would be a significant difference from how humans represent and reason about the same relations.

We have seen how symmetry has been considered a fundamental property of natural language, that helps making systematic efficient inferences about relations (18). Thus, the LLMs struggle in formal reasoning and spatial reasoning tasks can be imputed, at least partly, to their inability to represent symmetry.

It has been observed that the efficacy of linguistic communication among humans is significantly reliant on non-linguistic brain regions (6). This implies that language in isolation may be insufficiently informative, as it often relies on implicit cues and understandings (10).

Considering that the knowledge base of LLMs predominantly stems from textual data, there is a potential loss of information that humans typically derive through additional sensory inputs, which do not ne-

cessitate being stated explicitly. Thus, it is important to explore new ways in which these models can overcome the intrinsic loss of information that language brings.

Two main approaches have been used to solve this limitation of LLMs. One is *modularity*, which consist of building models with multiple specialised components that would support the language model in non-linguistic tasks (6). A different approach consists in making the language representation of these models to reflect human representation of concepts associated with words in a systematic and non-statistical-reliant way (24).

In the future, it would be useful to study more about how LLMs represent concepts that are expressed through language but that rely on other humans' abilities.

7 Conclusion

Our study focused on comparing the ways in which LLMs and humans represent symmetrical spatial relations. Initially, we have recognized the fundamental role of symmetry in natural language and its contribution to effective communication. Subsequently, our attention was directed towards the primary disparity between LLMs and humans in their representation of space and formal reasoning, emphasizing the potential factors contributing to this divergence.

In our experiment, we used BERT, an encoder that has proven to be able to represent linguistic knowledge rather than rely only on statistics regularities (14). The main question to address was how BERT represent symmetrical positional statements and whether it accounts for symmetry.

Spatial relations were thought of as transformation matrices altering vector entities, with inverse spatial relations corresponding to inverse matrices. We used a neural network to get these matrices from BERT and then used cosine similarity to compare the embeddings made by our method and BERT.

Our findings suggest a notable difference in BERT's ability to mirror human-like symmetry in representing symmetrical spatial relations.

We implement our code on Google Colab and make our code available on GitHub.[1]

References

[1] Ashish Vaswani, Ashish Vaswani, Noam Shazeer, Noam Shazeer, Niki Parmar, Niki Parmar, Jakob Uszkoreit, Jakob Uszkoreit, Llion Jones, Llion Jones, Aidan N. Gomez, Aidan N. Gomez, Łukasz Kaiser, Łukasz Kaiser, Illia Polosukhin, and Illia Polosukhin. Attention is All you Need. 30:5998–6008, June 2017. MAG ID: 2963403868.

[2] Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretič, and Samuel R. Bowman. Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 2877–2887, September 2019. doi: 10.18653/v1/d19-1286. URL https://arxiv.org/abs/1909.02597v2. ISBN: 9781950737901 Publisher: Association for Computational Linguistics.

[3] Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. The Better Your Syntax, the Better Your Semantics? Probing Pretrained Language Models for the English Comparative Correlative. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 10859–10882, October 2022. doi: 10.18653/v1/2022.emnlp-main.746. URL https://arxiv.org/abs/2210.13181v1. Publisher: Association for Computational Linguistics (ACL).

[4] Najoung Kim and Paul Smolensky. Testing for Grammatical Category Abstraction in Neural

[1] https://github.com/Thomasbush9/python_project_clc

Language Models. *Proceedings of the Society for Computation in Linguistics*, 4:60, 2021. doi: 10.7275/2nb8-ag59.

[5] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *EMNLP 2018 - 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Proceedings of the 1st Workshop*, pages 353–355, April 2018. doi: 10.18653/v1/w18-5446. URL https://arxiv.org/abs/1804.07461v3. ISBN: 9781948087711 Publisher: Association for Computational Linguistics (ACL).

[6] Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. January 2023. URL https://arxiv.org/abs/2301.06627v2.

[7] Brenden Lake and Marco Baroni. Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks. 2018. URL https://github.com/.

[8] Nils Nyberg, E. ´ Lé Onore Duvelle, Caswell Barry, and Hugo J. Spiers. Spatial goal coding in the hippocampal formation. doi: 10.1016/j.neuron.2021.12.012. URL https://doi.org/10.1016/j.neuron.2021.12.012.

[9] Chelsea Tanchip, Lei Yu, Aotao Xu, and Yang Xu. Findings of the Association for Computational Linguistics Inferring symmetry in natural language.

[10] N. J. Enfield. *Language vs. Reality: Why Language Is Good for Lawyers and Bad for Scientists*. MIT Press, March 2022. ISBN 978-0-262-04661-9. Google-Books-ID: nLZNEAAAQBAJ.

[11] Neil Burgess. Spatial memory: how egocentric and allocentric combine. doi: 10.1016/j.tics.2006.10.005. URL www.sciencedirect.com.

[12] George Lakoff and Mark Johnson. The Metaphorical Structure of the Human Conceptual System. *COGNITIVE SCIENCE*, 4:195–208, 1980.

[13] Alexandra O. Constantinescu, Jill X. O'Reilly, and Timothy E. J. Behrens. Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468, June 2016. doi: 10.1126/science.aaf0941. URL https://www-science-org.tilburguniversity.idm.oclc.org/doi/full/10.1126/science.aaf0941. Publisher: American Association for the Advancement of Science.

[14] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT Rediscovers the Classical NLP Pipeline. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 4593–4601, May 2019. doi: 10.18653/v1/p19-1452. URL https://arxiv.org/abs/1905.05950v2. ISBN: 9781950737482 Publisher: Association for Computational Linguistics (ACL).

[15] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience 2022 25:3*, 25(3):369–380, 2022. ISSN 1546-1726. doi: 10.1038/s41593-022-01026-4. Publisher: Nature Publishing Group.

[16] Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. *NAACL HLT 2018 - 2018 Confer-*

ence of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1:1195–1205, March 2018. doi: 10.18653/v1/n18-1108. URL https://arxiv.org/abs/1803.11138v1. ISBN: 9781948087278 Publisher: Association for Computational Linguistics (ACL).

[17] Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. What do RNN Language Models Learn about Filler-Gap Dependencies? *EMNLP 2018 - 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Proceedings of the 1st Workshop*, pages 211–221, August 2018. doi: 10.18653/v1/w18-5423. URL https://arxiv.org/abs/1809.00042v1. ISBN: 9781948087711 Publisher: Association for Computational Linguistics (ACL).

[18] Manasi Sharma. Exploring and Improving the Spatial Reasoning Abilities of Large Language Models. December 2023. URL https://arxiv.org/abs/2312.01054v1.

[19] Yuhan Ji and Song Gao. Evaluating the Effectiveness of Large Language Models in Representing Textual Descriptions of Geometry and Spatial Relations (Short Paper). Schloss-Dagstuhl - Leibniz Zentrum für Informatik, 2023. doi: 10.4230/LIPIcs.GIScience.2023.43. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.GIScience.2023.43.

[20] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity, November 2023. URL http://arxiv.org/abs/2302.04023. arXiv:2302.04023 [cs].

[21] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi Quoc, V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837, December 2022.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs].

[23] Marco Baroni and Roberto Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. (11):1183–1193, 2010. Publisher: Association for Computational Linguistics.

[24] Francisco E. De Sousa and Webber Vienna. Semantic Folding Theory And its Application in Semantic Fingerprinting. November 2015. URL https://arxiv.org/abs/1511.08855v2.