Reversing the NLP Pipeline: What Do LMs Have to Offer Linguistics?

Anonymous ACL submission

Abstract

Language models (LMs) have profoundly reshaped the field of natural language processing (NLP) in recent years. This paper presents the first investigation of the implications of LMs for linguistics, examining their role as a source of linguistic data. Despite their limitations, LMs have shown remarkable fluency and proficiency in modeling human language, suggesting their potential as linguistic artifacts. By proposing approaches to uncover these encoded linguistic structures, the paper explores how LMs provide opportunities for linguists. Embracing LMs as linguistic artifacts reopens the door for collaboration between NLP and linguistics, overcoming limitations of previous paradigms and providing new opportunities for the field of linguistics research.

1 Introduction

Language models (LMs), especially large language models (LLMs), have caused a revolution in the natural language processing domain. Replacing the previous linguistically-informed pipeline, they ushered in a new paradigm for computational processing of human language. While some may despair at the decreasing importance of linguistics in modern NLP, these developments provide an opportunity for uncovering knowledge about human language since, as shall be argued in this paper, language models represent a linguistic artefact in their own right.

Despite their shortcomings in real-world AI applications, such as hallucinations (OpenAI, 2024; Dash et al., 2023), it is clear that LMs can accurately model human language, as evidenced by their surprisingly fluent conversational ability. Crucially they can generate sentences never before seen in training. The fluency of their text output has even led some to claim that ChatGPT has passed the Turing test (Biever, 2023), being indistinguishable from a human in short conversations in a text format.

One of the main aims of (synchronic) linguistics as a field is "to account for the potentially infinite set of well-formed sentences in all languages" (Rastall, 2010), that is, to uncover the structure and productive rules of human language. And it just so happens that language models, albeit covertly, encode at least a large majority of these structures, if not all, as evidenced by their conversational ability.[1] This provides an as of yet unused opportunity to extract the embedded language structures for the purpose of furthering linguistic knowledge.

In this paper I would therefore like to show the great potential of language model analysis for the field of linguistics and to propose some methods of extracting linguistic knowledge from LMs, thus showing that the benefits of collaboration between linguistics and NLP are not a one-way street.

2 Some current computational approaches to linguistics

Computational approaches to linguistics are nothing new, and this section aims to outline some of the relevant previous work in this field.

2.1 Corpus linguistics

The advent of the computer era in the 1960s brought with it the appearance of digital text corpora. This made it possible to study what Firth calls "attested language" (Anderman and Rogers, 2007) on a large scale for the first time. English corpora such as the Brown Corpus (Francis and Kucera, 1979) or the British National Corpus (Leech, 1992), consisting of millions of tokens from texts of different domains, were a valuable tool for theoretical and computational linguists alike. This made it

[1] This is to say nothing about the question of their consciousness, cf. Searle's Chinese Room Argument (Searle, 1980).

possible to empirically test claims that are otherwise hard to validate reliably through introspection (Hunston, 2022). It also made linguistics more reproducible and objective, abstracting away from the linguistic bias of singular speakers or groups of researchers.

Written and spoken corpora have contributed enormously to linguistics, and despite Chomsky's misgivings their use has now become commonplace in the field. An adequate discussion of their contribution to the field would go beyond the scope of this paper, but some examples of using corpus data to show hypotheses are O'Keeffe (2007) from the ESOL domain, Kesebir and Kesebir (2012) from psychology or Meurers (2005) in syntax. Historical linguistics also relies heavily on corpora, since it is of course impossible to ask a native speaker in many cases.

2.2 Word embeddings

Arguably a precursor to modern LMs, building on the corpora discussed in 2.1, word embeddings (Mikolov et al., 2013; Pennington et al., 2014) aim to represent words (or tokens) as a real vector $\mathbf{v} \in \mathbb{R}^d$. These are usually calculated so that words that often appear together in training corpora have a high similarity. This real-world encoding of Firth's famous hypothesis that "a word is characterized by the company it keeps" (Firth, 1957) opened the door to serious empirical linguistic research with the formerly purely theoretical distributional hypothesis of semantics.

Using vector word representations embedded in time, Hamilton et al. (2016) showed how diachronic semantic changes in English terms could be rediscovered through comparing their embeddings rooted in time. Their experiments also empirically confirmed the hypothesis from diachronic linguistics that polysemous words change at faster rates.

Other examples are Basirat and Tang (2018), who were also able to convincingly predict grammatical features of nouns in Swedish such as the common/proper noun distinction and grammatical gender using word embeddings, and, on the sociolinguistic front, Caliskan et al. (2022), who demonstrated how gender bias pervases GloVe word-embeddings trained on internet corpora.

3 LMs in linguistics

Currently the two main approaches used to develop and validate linguistic hypotheses are through corpora and through introspection, the former being championed by empiricists and the latter by the Chomskyan rationalist tradition (McEnery and Wilson, 2001). It is clear that corpora, including those used to train LLMs, can only contain a fraction of the famously infinite set of possible grammatical sentences in a language, and this has led Chomsky to decry corpus linguistics as seeking to model language *performance* rather than *competence* (McEnery and Wilson, 2001). Native speaker introspection, on the other hand, while able to judge the grammaticality of any sentence, is clearly highly subjective and biased. Language models, however, are productive and are able to generalize across their corpora to produce, with some sophistication, sentences not seen before in training, thus blurring the line between the traditional Chomskyan distinction between competence and performance.

3.1 History and design of LMs

In order to analyze language models as a linguistic object, it is necessary to look at what exactly an LM is and where it has come from.

Statistical language modeling has its origins in the 1990s with models such as the *n*-gram model (Brown et al., 1992), which uses the chain rule to calculate the probability of the next token w_i based off of the previous *n* tokens using the following formula:

$$P(w_i) = P(w_i | w_{i-n+1}, ..., w_{i-1}) \quad (1)$$

In this case corpora were used to train these statistical models and thus calculate the *n*-gram probabilities which are used during generation.

The refinement of neural networks, however, lead to the first neural training of vector representations of words, as described in 2.2, in order to apply the geometric concept of (cosine) similarity to the semantic space. This was further refined by the idea of contextual word embeddings (Peters et al., 2018; Devlin et al., 2019), which allows for polysemy in word tokens. This contextualization allows the embeddings to take word order into account and thus to be able to model sentences much more accurately. BERT (Devlin et al., 2019) uses an attention mechanism to embed information from other tokens in a sentence in a word representation and was a breakthrough in the world of NLP, ushering in the age of pre-trained language models (PLMs). Their popularity lies in the fact that PLMs, which have been pre-trained on large text corpora, can subsequently be fine-tuned on many downstream NLP tasks such as sentiment classification or question-answer systems, achieving state of the art results in many of these tasks.

Increasing amounts of data and more efficient hardware led to the development of large language models (LLMs) (Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023), consisting usually of at least 1 billion parameters. Together with improvements on traditional NLP benchmarks, these models seem to exhibit other "emergent abilities" (Wei et al., 2022) such as logical reasoning or performing simple arithmetic,[2] also including tasks more relevant to linguistics such as POS tagging (Chopra, 2024).

Therefore LLMs, which are usually trained on such trivial tasks as minimizing their error on next word prediction on the training corpus, end up being able to produce fluent human-like text and can be used for a variety of tasks.

3.2 LMs as a linguistic artefact

The performance of LMs, particularly LLMs, implies that the linguistic structures underlying the written language they were trained on are stored, albeit implicitly, within the parameters of the model. This fact alone warrants their closer analysis by linguists: the generative capabilities of LMs combined with their lack of personal bias provides a third way in the current methodological opposition between rationalists and empiricists. Indeed ChatGPT can even be asked about the acceptability of a particular sentence (see 4.3). The black-box nature of neural networks (i.e. the arbitrary nature of their parameters outside of the whole parameter system) makes the task of examining their inner workings more complex, however in the following section I will propose some methods to extract linguistic knowledge from LMs.

4 Methods for extraction of linguistic knowledge from LMs

The use of LMs in linguistics research can be grouped into two main approaches: internal and external probing. Internal probing aims to find structure in the model parameters through mathematical or statistical techniques such as dimension reduction or clustering. One notable example of this is Tenney et al. (2019a), who used edge probing (Tenney et al., 2019b) to show how the classic NLP pipeline (starting with POS tagging and parsing, and ending with semantic roles and coreference) can be found in the layers of BERT.

External probing, on the other hand, utilizes the model's linguistic ability in order to analyze its output from a linguistic perspective, such as using it for annotation, to generate data or to query the acceptability of certain sentences.

4.1 Latent space analysis

The latent space of an LM, i.e. the embedding space $\mathbb{R}^d$, is a d-dimensional space where the vector representations of components of the natural language input reside. Analysis of this space through internal probing can be fruitful, leading to a deeper understanding of how LMs work the way they do and, by extension, how the language system itself is structured.

One minimal example of the use of this is the following:[3] by fine-tuning a BERT model on an English dataset annotated for one of five verbal aspect classes, it becomes possible to examine the embedding space used by the model (for more details see B). Figure 1 provides a visualization of the `[CLS]` token embedding of verb-sentence pairs in the training set, together with their aspect label. It is clear how this could be useful to, for example, empirically motivate the hypothesis that habituals inhabit an area of the semantic aspect space between activities and states, which also intuitively makes sense.

4.2 LMs for annotation

Among the professed "emergent capabilities" of LLMs (Wei et al., 2022) is a capacity for logical reasoning. While not perfect, LLMs do seem to have human-like abilities in many areas, which can be utilized for annotation of linguistic data. While of course to be used with caution, the performance of LLMs as annotators has been shown to rival or even surpass that of human annotators in some cases (Gilardi et al., 2023). An example use-case could be POS tagging of a large corpus in order to study the relative frequency of different syntactic structures, thus overcoming the difficulty of finding

[2] Albeit with imperfect results (Liu et al., 2023).

[3] Code available here.

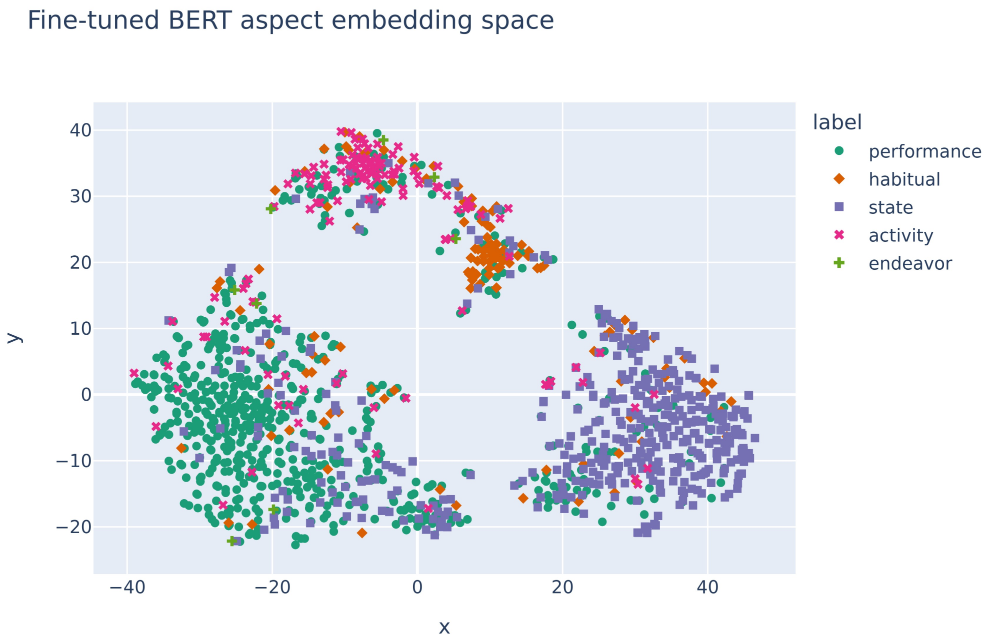

Figure 1: `[CLS]` embedding space of a BERT model fine-tuned on English verbs annotated for aspect in context, reduced to 2 dimensions by t-SNE.

and funding expert human annotators with systems capable of near-human performance (Bohnet et al., 2018).

This area also opens up possibilities for low-resource languages, where finding annotators may be more difficult, or where data is sparse. For example, Kholodna et al. (2024) show how LLMs can be used for named entity recognition (NER) in African languages such as isiZulu and Bambara, reaching a comparable (or better) performance compared to a human annotator.

4.3 LMs for acceptability analysis

A special case of using LMs for annotation is using them for acceptability analysis, a key feature of linguistic research. While most studies in linguistics use introspection to determine the acceptability of a certain utterance (since, as already mentioned, corpora are finite), this is a highly subjective and, in essence, unscientific process due to its lack of objectivity and observability.

This problem can be solved by querying LLMs on the acceptability of a sentence (see A for an example) or using an LM fine-tuned on a dataset such as CoLA (Warstadt et al., 2018), a dataset consisting of 10,657 sentences in English labeled as grammatical or ungrammatical taken from published linguistics literature. The sentences were reannotated by five linguistics students leading to an average annotator agreement of 86.2% with the original judgement in the paper, signifying the problem with introspection as a source of linguistic data. Taking this as a baseline, this has been surpassed by several neural network systems (Cherniavskii et al., 2022; Proskurina et al., 2023; Sileo, 2023), showing their suitability for such tasks. Using a language model for acceptability annotation increases the objectivity and reproducibility of the results, without sacrificing their validity.

5 Conclusion

In conclusion, LMs are a linguistic artefact and should be seen as such by the linguistics community. Acceptance of this assumption opens up a world of possibilities for linguistics research and could help overcome some of the drawbacks of previous paradigms.

The techniques presented in this paper serve as some examples for extracting linguistic knowledge from LMs, and I hope that this paper will inspire further interdisciplinary research leading to the development and refinement of more sophisticated methods.

6 Limitations

In light of the current hype surrounding deep learning it is important to highlight the limitations of such techniques and what language models *cannot* do.

It is well-known that training the LMs discussed in this paper requires a large amount of data and computing resources. While pre-trained models mean that LMs trained on relatively large amounts of data are now available for general use, for less well-resourced languages this is a problem, and their performance suffers drastically. While this does not rule out the use of LMs on such languages (see (Kholodna et al., 2024)), it certainly limits the applicability of some of the uses highlighted in this article, such as for annotation or acceptability analysis.

Furthermore, it must also be noted that LMs take on any biases present in the training data, meaning the language they approximate should be treated with caution. Examining these biases, as has often been done before, can, however, be an area of study in its own right and can produce valuable data for sociolinguistics. However, it must also be noted that it cannot be guaranteed that characteristics of a model's latent space can be transferred to a more general linguistic space, since human linguistic competence and LM competence differ in some aspects. Further research is therefore needed in this area.

Finally, since LMs are trained to minimize error on one variety of a language, they are less well-suited to study linguistic variation, whether geographical or temporal. This makes their use less suitable for languages without an accepted standard variant, such as Swiss German. In these cases, however, an approach using word embeddings such as Hamilton et al. (2016) could still be useful.

References

Gunilla Anderman and Margaret Rogers. 2007. *Chapter 1. The Linguist and the Translator*, pages 5–17. Multilingual Matters, Bristol, Blue Ridge Summit.

Ali Basirat and Marc Tang. 2018. Lexical and Morpho-syntactic Features in Word Embeddings - A Case Study of Nouns in Swedish. In *Special Session on Natural Language Processing in Artificial Intelligence*, pages 663–674, Funchal, France. SCITEPRESS - Science and Technology Publications.

Celeste Biever. 2023. Chatgpt broke the turing test — the race is on for new ways to assess ai. *Nature*.

Bernd Bohnet, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, and Joshua Maynez. 2018. Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2642–2652, Melbourne, Australia. Association for Computational Linguistics.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. 1992. Class-based *n*-gram models of natural language. *Computational Linguistics*, 18(4):467–480.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R. Banaji. 2022. Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22. ACM.

Daniel Chen, Martha Palmer, and Meagan Vigus. 2021. AutoAspect: Automatic annotation of tense and aspect for uniform meaning representations. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 36–45, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniil Cherniavskii, Eduard Tulchinskii, Vladislav Mikhailov, Irina Proskurina, Laida Kushnareva, Ekaterina Artemova, Serguei Barannikov, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2022. Acceptability judgements via examining the topology of attention maps. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics.

Sahil Chopra. 2024. Zero resource cross-lingual part of speech tagging.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben

Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Debadutta Dash, Rahul Thapa, Juan M. Banda, Akshay Swaminathan, Morgan Cheatham, Mehr Kashyap, Nikesh Kotecha, Jonathan H. Chen, Saurabh Gombar, Lance Downing, Rachel Pedreira, Ethan Goh, Angel Arnaout, Garret Kenn Morris, Honor Magon, Matthew P Lungren, Eric Horvitz, and Nigam H. Shah. 2023. Evaluation of gpt-3.5 and gpt-4 for supporting real-world information needs in healthcare delivery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

J.R. Firth. 1957. *A Synopsis of Linguistic Theory, 1930-1955*.

W. N. Francis and H. Kucera. 1979. Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Jens Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos-Yopán, and Nianwen Xue. 2021. Designing a uniform meaning representation for natural language processing. *KI - Künstliche Intelligenz*, 35.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Susan Hunston. 2022. *Corpora in Applied Linguistics*, 2 edition. Cambridge Applied Linguistics. Cambridge University Press.

Pelin Kesebir and Selin Kesebir. 2012. The cultural salience of moral character and virtue declined in twentieth century america. *The Journal of Positive Psychology*, 7(6):471–480.

Nataliia Kholodna, Sahib Julka, Mohammad Khodadadi, Muhammed Nurullah Gumus, and Michael Granitzer. 2024. Llms in the loop: Leveraging large language model annotations for active learning in low-resource languages.

Natalia Klyueva and Ondřej Bojar. 2008. UMC 0.1: Czech-Russian-English Multilingual Corpus. In *Proceedings of International Conference Corpus Linguistics*, pages 188–195.

Geoffrey Leech. 1992. 100 million words of english:the british national corpus (bnc). *Second Language Research*, 28:1–13.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4.

Tony McEnery and Andrew Wilson. 2001. *Corpus Linguistics: An Introduction*. Edinburgh University Press.

W. Detmar Meurers. 2005. On the use of electronic corpora for theoretical linguistics: Case studies from the syntax of german. *Lingua*, 115(11):1619–1639. Data in Theoretical Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

OpenAI. 2024. Gpt-4 technical report.

Anne O'Keeffe. 2007. *O'Keeffe, A., McCarthy, M. J. and R. A. Carter (2007) From Corpus to Classroom: language use and language teaching. Cambridge: Cambridge University Press.*

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Irina Proskurina, Ekaterina Artemova, and Irina Piontkovskaya. 2023. Can BERT eat RuCoLA? topological data analysis to explain. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 123–137, Dubrovnik, Croatia. Association for Computational Linguistics.

Paul Rastall. 2010. Three goals of linguistic explanation. *La Linguistique*, 46(1):3–18.

John R. Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424.

Damien Sileo. 2023. tasksource: A dataset harmonization framework for streamlined nlp multi-task learning and evaluation.

Olli O. Silvennoinen. 2023. Not just contrastive: Constructions with negated restrictives in english. *Journal of English Linguistics*, 51(4):346–374.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. Bert rediscovers the classical nlp pipeline.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Zeno Vendler. 1957. Verbs and times. *The Philosophical Review*, 66(2):143–160.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

A ChatGPT and linguistic acceptability

The following is a script from an interaction with ChatGPT 3.5 on 20/03/2023:

Prompt: *Is the following sentence grammatically and syntactically acceptable: "He not merely likes her, he loves her."*

ChatGPT: *Yes, the sentence "He not merely likes her, he loves her." is both grammatically and syntactically acceptable. It follows a parallel structure, presenting two related ideas in a balanced way. However, it's worth noting that the phrase "not merely" is a bit formal and might sound slightly old-fashioned in casual conversation. A more common way to express the same sentiment would be: "He doesn't just like her, he loves her."*

While Silvennoinen (2023) argues that the above sentence is unacceptable, a search on Google n-grams reveals that examples with the phrase "not merely" followed by a finite verb construction can be found, especially in earlier texts, validating ChatGPT's response.

However, ChatGPT agreed with the author on the inacceptability of the following case:

Prompt: *Is the following sentence acceptable: "He not only likes her."*

ChatGPT: *The sentence "He not only likes her" is grammatically incomplete. It lacks the second part of the comparison, which is necessary for clarity and grammatical correctness. You need to follow up "not only" with "but also" or some other comparative phrase to complete the structure. For example:*

"He not only likes her, but he also admires her."

"He not only likes her, but he also respects her."

Adding the "but also" or similar comparative phrase makes the sentence grammatically acceptable and conveys a clear comparison.

B Experiment details

In order to calculate the embeddings shown in figure 1, a `BERT-base-uncased` model (110M parameters) was fine-tuned for 5 epochs on a selection of 4008 verbs together with the sentence they appear in from the English part of the UMC003 trilingual dataset (Klyueva and Bojar, 2008). The verb-sentence pairs were annotated by a fine-tuned Llama 2 model (Touvron et al., 2023), outputting one of the 5 aspect classes in the Uniform Meaning Representation (UMR) framework (Gysel et al., 2021). The reason for the choice of this aspect classification was the availability of a small amount of training data for fine-tuning the Llama 2 model.

Gysel et al. (2021) defines the UMR aspect classes as follows:

- State - The State value corresponds to stative events in Vendler (1957); no change occurs during the event. It also includes predicate nominals (*be a doctor*), predicate locations (*be in the forest*), and thetic (presentational) possession (*have a cat*).

- Habitual - The Habitual value is annotated on events that occur regularly in the past or present.

- Activity - The Activity value indicates an event has not necessarily ended and may be ongoing at Document Creation Time (DCT).

- Endeavour - Endeavor is used for processes that end without reaching completion (i.e., termination).

- Performance - Performance is used for processes that reach a completed result state.

For a more in-depth description of the UMR aspect classes, please see Gysel et al. (2021); Chen et al. (2021).

The fine-tuned BERT model was then given datapoints from the test set and the embedding of the `[CLS]` classifier token was reduced to 2 dimensions by t-SNE (van der Maaten and Hinton, 2008). This was then plotted in figure 1 together with the aspect label to show the topology of the LM's aspect space.