Structuring the unstructured: an LLM-guided transition

Alessandro De Bellis[1]

[1]*Politecnico di Bari, Bari, Italy*

Abstract

The surge of interest in Large Language Models (LLM) has reached unprecedented levels of magnitude in the last year. Following the success of ChatGPT, a powerful conversational system powered by an LLM, both specialized and non-specialized users have started to leverage the efficacy of these systems. If this trend is to be maintained, it is reasonable to predict an ever more ubiquitous role for LLMs in our daily life. Nonetheless, the widespread adoption of LLMs in various applicative fields remains an ongoing challenge. In order to achieve competitive levels of performance in specialized tasks, LLMs can require complex fine-tuning procedures and high-quality data. Furthermore, LLMs are scarcely interpretable and often viewed as "black boxes". This proposal aims to lessen the gray areas around the subject of LLMs and shed some light on which kind of information they store internally. Specifically, the goal is to assess the ability of LLMs to model external semantic knowledge about concepts encoded from unstructured text, by means of their latent representations. The proposal aims to explore existing patterns in the latent space that convey explicit information about taxonomical information and relational connections between concepts, and that can therefore reflect the knowledge encoded by a Knowledge Graph. The resulting knowledge could be exploited to enable complex downstream tasks leveraging factual knowledge and aimed to deduce new structured information, possibly in zero-shot or few-shot settings.

Keywords

Large Language Models, Knowledge Graphs, Information Extraction

1. Introduction

Extraction of structured data from unstructured data is a complex task that combines various levels of natural language understanding. Although Language Models (LMs) are proven to model effectively complex syntax dependencies and semantic relationships in sentences, some tasks require more deep language understanding capabilities that focus on pragmatics and external knowledge. To this purpose, numerous approaches driven by Large Language Models (LLMs) have been proposed, most of which involve conditioning on fine-tuning tasks, typically in distant-supervised settings. However, the complexity of fine-tuning may be a concern in some applications. In addition, fine-tuning for a specific task requires the availability of large volumes of high-quality annotated data, which is not always possible. Even when feasible, fine-tuning results in hyperspecialized pipelines that cannot translate to new tasks or domains. Lastly, those approaches are typically "black-box" as they are based on the formulation of an end-to-end task. In most cases, the embedding space is never investigated.

In addition, it is worth noting a growing common concern about the lack of control over what these models learn, since it may lead to unpredictable results and toxic or biased content. Interpretability for LLMs is still largely an unsolved problem that asks for immediate attention.

Previous works [1] have underlined how large, pretrained Language Models are able to acquire vast amounts of knowledge that transcend mere grammar and general patterns. Nonetheless, the exploration of the underlying learning mechanisms of these models, which pertains to the understanding of what information they actually acquire, is a relatively new research direction yet to be fully addressed. This work aims to interpret the inner knowledge of LLMs as modeled by their intermediate representations. The moving hypothesis is that, by observing large volumes of heterogeneous textual data, LLMs are able to capture vast amounts of factual knowledge (i.e. verifiable information about the world, such as people, locations, events, etc.). Understanding where this knowledge is stored, how to access it and interpret it has the potential to drive progress in this research field and enable the application of LLMs in various semantic web applications. In particular, this proposal aims to explore the similarities between LLM latent spaces and Knowledge Graphs (KGs). By analyzing the latent spaces of LLM, we can identify patterns that align them with the structure of KGs. In addition, this knowledge could be exploited to extract new knowledge from the LLMs.

In a prior study (Anelli et al. [2]), we investigated the presence of significant regions within the latent space of BERT [3]. The identification of distinct boundaries within the latent spaces of LLMs, aligning with the taxonomical organization of Knowledge Graphs (KGs), offers the potential for leveraging these boundaries to establish connections between text and real-world concepts. This has the potential to enable ontology-driven semantic parsing and tagging applications on unstructured text, even in scenarios with limited data or without prior conditioning on text. In addition, KGs could be exploited to find relational patterns in LLMs latent spaces, enabling tasks such as Knowledge Graph Completion (KGC) and Link Prediction (LP) from unstructured text.

The remainder of the paper is structured as follows: Section 2 illustrates the theoretical notions at the basis of this proposal, as well as a literature review. Section 3 provides a more in-depth discussion of the preliminary hypothesis and research questions. Section 4 illustrates the expected methodologies and evaluation strategies required for the research path. Section 5 concludes the paper with final remarks.

2. Background

2.1. Large Language Models

The problem of Language Modeling consists in assigning a probability to a sequence of words. Capturing the inner nature of natural language remains a heavily data-dependent problem. For this reason, the paradigm of *transfer learning* was investigated. Recent approaches for pretraining rely on a self-supervised pretext task, namely Masked Language Modeling (MLM), with the goal of completing artificially masked parts of a sentence. Through this general task where the model learns how to fill the gaps, it learns general linguistic constructs, syntax substructures, patterns as well as semantic characteristics, both at sentence and at discourse level. This general knowledge, condensed into a latent space, can be then exploited to perform

more specific downstream tasks by means of fine-tuning.

The introduction of the Transformer [4] architecture marked a milestone in the history of language modeling. Transformers surpass the shortcomings of LSTM architectures being able to process longer sequences of words simultaneously and capturing long-range dependencies. One of the first evidences of this emerging paradigm was BERT [3], an encoder-only Pretrained Language Model (PLM) trained on a MLM task.

The recently coined expression "Large Language Model" broadly refers to a class of PLMs that are trained on large volumes of textual data, possibly including several billions of sentences, and comprise a significant amount of internal parameters. GPT [5], introduced by OpenAI, employs a Transformer decoder architecture which is pretrained on an unsupervised unidirectional language modeling task. Subsequently, GPT-2 [6] demonstrated that a LLM pretrained on a sufficiently large dataset with enough diversity is enough to perform various zero-shot tasks with no need for supervision. GPT-3 [1] proved that scaling up the pretraining task and the number of parameters (175 billions) allows for surprising few-shot performances, reaching competitive levels even against fine-tuned baselines. Later on, OpenAI reached its latest milestone with GPT-4 [7], a multimodal model capable of accepting text as well as images as input.

2.2. Knowledge Graphs

Knowledge Graphs (KGs) are a form of structured data representation for relational information about real-world objects (i.e. entities). A KG can also be enhanced with additional metadata and properties, both structured and unstructured, about the entities. A schema, defined by means of an ontology, enforces a specific taxonomy for the entities and relationships in a KG. Formally, a KG can be defined as a triple $G = \{E, R, F\}$, where E is a set of entities, R is a set of relations and F is a set of facts. A fact $f \in F$ is a triple (h, r, t) where $h, t \in E$ and $r \in R$.

KGs allow an agent for easy access and retrieval of explicit information, as well as the application of reasoning and logical inference techniques for implicit information extraction. In addition, KGs paved the way for the development of techniques aimed at the automatic inference of new knowledge. Knowledge Representation Learning (KRL) identifies an emerging field of study whose goal is to map the discrete concepts in a KG to dense, low-dimensional vectors (i.e. embeddings) in a semantic space. Embedded representations address the challenges of computational and memory complexity by having a lower dimensionality. Furthermore, by mapping discrete concepts to a continuous vector space, it is possible to alleviate the effects of data sparsity often present in common KGs. These embedded representations can be exploited in various machine learning end tasks, such as Knowledge Graph Completion (KGC), whose goal is to fill incomplete triples inferring the missing components. Furthermore, embeddings capture semantic similarities between concepts, providing practical value in Node Classification, which involves mapping entities in a KG to their corresponding categories. Information Retrieval (IR) subtasks can also benefit from the robustness of embeddings with respect to data sparsity and the lower computational constraints with respect to traditional approaches.

Previous works [8] proposed the integration of textual information into knowledge representation learning (KRL), aiming to create unified representational spaces that encompass both text and KGs. The intermixture of text and discrete knowledge can be beneficial for Semantic Annotation [9], which aims to identify key concepts in text and link them to concepts in KGs.

2.3. Related Work

Other works attempted to associate a human interpretable meaning to the behavior of LMs at a more deep inner level (i.e. hidden activations or attention weights). Works such as the one of Bolukbasi et al. [10] attempt to measure the impact of various inputs on the neural activations of the BERT [3] model, showing that this model exhibits recognizable activation patterns for similar concepts and proving the existence of meaningful global directions in the BERT space. However, the authors also found that the activation interpretation varied to some extent with respect to the dataset of choice. Manning et al. [11] analyzed the structural properties of the BERT latent space, proving through structural probing that a parse tree can be reconstructed from the hidden representations of the architecture. This work demonstrates that latent representations coming from a self-supervised approach such as Masked Language Modeling indirectly acquire implicit information about syntactic intra-sentence relations.

Petroni et al. [12] assessed the ability of several popular pretrained language models to act as queryable "off-the-shelf" factual knowledge bases. The underlying assumption is that the standard unidirectional or bidirectional language modeling problem formulation is strictly related to the task of cloze-style question answering. However, the latent representations are not explored in this work, as the task is based on leaving the end-to-end architecture unchanged while solving the task at prompt-level. Prompt-based approaches might fail at retrieving relevant knowledge from an LM based on the prompt choice. As noted by Jiang et al. [13], existing experimental evidence about the ability of LMs to capture factual knowledge might only represent a lower bound of what these architectures actually know, since we do not possess accurate and proven ways to probe this knowledge.

Several works have attempted to extract factual knowledge from LMs embedding spaces. Yao et al. [14] attempt to employ LMs in the context of knowledge graph completion: they represent KG triples as textual sentences and extract a representation space through BERT; this representation space is then conditioned on a plausibility function through fine-tuning. In a similar fashion, [15] proposed an approach for KGC based on the fine-tuning of GPT-2. These approaches prove to be effective in surpassing traditional knowledge graph embedding ones, for their ability to model rich semantic information instead of simple structure. However, they rely only on fine-tuning end-to-end architectures, while the existing semantic information about factual knowledge in the pre-trained space is not investigated. Contrarily, this proposal aims to investigate whether various pretrained LLM latent spaces can directly incorporate this knowledge as well as preserving the desired language understanding capabilities that come with a task-agnostic pretraining procedure.

An attempt to interpret what the BERT [3] Language Model learns through its pretraining task was made in our previous work (Anelli et al. [2]). This work was moved by the assumption that pre-trained BERT latent space encapsulates semantics about real-world objects that can be leveraged to construct facts via a link prediction procedure. The assumption was then proven by evaluating against the FB15K-237 [16] benchmark dataset. Following this intuition, an entity classification task was then performed to assess the separability of the same space with regard to the taxonomy enforced by the Freebase ontology.

3. Research Proposal

The goal of this research proposal is to answer the following questions:

- **Q1**: do LLM latent spaces exhibit patterns of semantic grounding in structured open-domain knowledge bases?

  - **Q1.1**: is factual and ontological knowledge encoded in an LLM latent space? Is it possible to extract it? Does the extracted knowledge match with the KG ground truth? How many entities and what properties can be extracted? How do different state-of-the-art LLMs compare in this regard?
  - **Q1.2**: at which stage or by which parameters of the LLM architecture is this semantic information mainly encoded? Are certain hidden parameters more suited to specific facts/domains?
  - **Q1.3**: can we associate an explicit interpretation to geometrical characteristics of LLM latent representations (e.g. regions, directions, distances, linear transformations, separability) that links them to ontological and relational information?
  - **Q1.4**: could this knowledge about the latent representations be exploited to perform a wide variety of IE-related downstream tasks, such as Knowledge Graph Completion and Semantic Tagging, especially in few-shot settings? If so, could they yield state-of-the-art results?
  - **Q1.5 (related question)**: what role does context play in the semantics encoded by the LLM latent representation?
  - **Q1.6 (related question)**: does this semantic awareness transfer well to closed domains as well without the need for fine-tuning? Are pretrained LLMs biased towards a specific niche domain or a set of properties?

- **Q2**: can we induce a semantic representation space from text that models the ontological properties and relational information of a structured knowledge base, while preserving the desirable linguistic understanding capabilities of LLMs?

  - **Q2.1**: could this be done by exploiting existing state-of-the-art architectures through fine tuning?
  - **Q2.2**: would a new architecture or training procedure be more suited for the purpose?
  - **Q2.3**: could a more appropriate loss function or training procedure be formulated for this purpose?

The underlying goal of this proposal is to enable LLMs as general knowledge-aware *feature extractors* to obtain either "universal" or domain-specific knowledge representation of concepts from text. These representations should be highly interpretable and task-agnostic, such that they could accommodate for a variety of Information Extraction tasks "*off-the-shelf*", acting only at the discriminative level of the architecture and with little to no need for annotated data.

Lastly, the research work will put significant emphasis on ensuring the reproducibility of the results, with the aim of facilitating access for the community and enabling potential improvements.

4. Methodologies and Evaluation

The first part of the proposed research will be devoted to the extraction of patterns of semantic grounding of LLMs latent representations in knowledge bases. A latent space offers an abstract representation of discrete concepts by means of real-valued numerical vectors. Post-hoc techniques for the analysis and the interpretation of latent space coming from deep-learning models remain under active research. Some recent approaches [17] rely on a linear mapping operation from a latent space into a pre-defined semantic space with explicit interpretability.

An important semantic grounding aspect of latent representation pertains to how their distributional characteristics can reflect a human-understandable taxonomy. Part of this research will try to assess the existence of explicit, separable semantic regions in LLMs latent spaces. Linear Probing was originally introduced by Alain et al. [18] as a tool to investigate the role of different hidden layers of deep neural networks. A linear "probe" is a linear classifier trained on top of an intermediate layer to predict a label. The measured accuracy of said classifier allows to assess the semantics of each layer and the degree of separability for the resulting representation. Probing-based unsupervised techniques, such as clustering, acting at intermediate layers of LLM architectures, might be adopted to extract patterns of separability existing in a LM latent space enforced by taxonomical categorization and ontology.

In a second phase, emerging patterns of relational semantics in latent spaces will be explored. As proven by Bolukbasi et al. [10], global directions in LM embedding spaces can encode different meanings and concepts. Voynov et al. [19] propose an unsupervised approach to discover interpretable meaning for specific directions in a generative model latent space; this approach was applied in the context of image generation to prove the existence of a relationship between vector linear transformation and human-interpretable image transformations. Similar approaches could be adopted to explain the semantics of local and global directions in a LLM latent space.

Lastly, the research will focus on transferring the knowledge about the semantics of latent representations to downstream Information Extraction (IE) tasks. Specifically, Knowledge Graph Completion (KGC) represents a clear benchmark use case for our findings, since recent studies have started to investigate LLM-augmented KGC techniques to tackle issues deriving from KG sparsity, as well as providing a way to handle out-of-vocabulary (OOV) concepts.

Existing evaluation strategies for most KGC techniques leverage ranking-based procedures. Results can be hard to interpret since these metrics are often based on a quantity over quality principle, as they establish whether or not a model is able to predict large amounts of facts correctly, while they do not take into account how "hard" those predictions can be. Some facts are particularly trivial to classify by logical inference (e.g. inverse relations). In addition, benchmark datasets for this task are often constructed without particular constraints, resulting in datasets that are often biased towards a niche domain or contain a large portion of trivial facts. In order to provide extensive coverage for various properties and entities in KGs, the extraction of a new benchmark dataset might be required.

5. Conclusions and Future Work

The research proposal presented in this paper starts from the hypothesis that LLMs are inherently latent knowledge bases with interpretable semantics. By virtue of this assumption, the research aims to determine (i) whether LLM latent representation encode semantics about world factual knowledge, (ii) where this information is encoded and (iii) whether this information can be exploited in real applicative scenarios. The resulting outcome could have the potential to offer a deeper understanding of the underlying mechanisms of these models. Moreover, identifying semantics in the LLMs' latent space could broaden the applicative scope of LLMs, paving the way to a deeper integration of LLMs with semantic technologies.

Acknowledgments

References

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, CoRR abs/2005.14165 (2020). URL: https://arxiv.org/abs/2005.14165. arXiv:2005.14165.

[2] V. W. Anelli, G. M. Biancofiore, A. De Bellis, T. Di Noia, E. Di Sciascio, Interpretability of bert latent space through knowledge graphs, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 3806–3810. URL: https://doi.org/10.1145/3511808.3557617. doi:10.1145/3511808.3557617.

[3] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.

[5] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training (2018).

[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners (2019).

[7] OpenAI, Gpt-4 technical report, 2023. `arXiv:2303.08774`.

[8] J. Wu, R. Xie, Z. Liu, M. Sun, Knowledge representation via joint learning of sequential text and knowledge graphs, CoRR abs/1609.07075 (2016). URL: http://arxiv.org/abs/1609.07075. `arXiv:1609.07075`.

[9] L. Wang, J. Lu, G. Zhou, H. Pan, T. Zhu, N. Huang, P. He, C. De Maio, Representation learning method with semantic propagation on text-augmented knowledge graphs, Intell. Neuroscience 2022 (2022). URL: https://doi.org/10.1155/2022/1438047. doi:`10.1155/2022/1438047`.

[10] T. Bolukbasi, A. Pearce, A. Yuan, A. Coenen, E. Reif, F. Viégas, M. Wattenberg, An interpretability illusion for bert, 2021. `arXiv:2104.07143`.

[11] C. D. Manning, K. Clark, J. Hewitt, U. Khandelwal, O. Levy, Emergent linguistic structure in artificial neural networks trained by self-supervision, Proceedings of the National Academy of Sciences 117 (2020) 30046 − 30054.

[12] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, A. Miller, Language models as knowledge bases?, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2463–2473. URL: https://aclanthology.org/D19-1250. doi:`10.18653/v1/D19-1250`.

[13] Z. Jiang, F. F. Xu, J. Araki, G. Neubig, How can we know what language models know?, Transactions of the Association for Computational Linguistics 8 (2020) 423–438. URL: https://aclanthology.org/2020.tacl-1.28. doi:`10.1162/tacl_a_00324`.

[14] L. Yao, C. Mao, Y. Luo, Kg-bert: Bert for knowledge graph completion, ArXiv abs/1909.03193 (2019).

[15] R. Biswas, R. Sofronova, M. Alam, H. Sack, Contextual language models for knowledge graph completion, in: MLSMKG@PKDD/ECML, 2021.

[16] K. Toutanova, D. Chen, Observed versus latent features for knowledge base and text inference, in: Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality, Association for Computational Linguistics, Beijing, China, 2015, pp. 57−66. URL: https://aclanthology.org/W15-4007. doi:`10.18653/v1/W15-4007`.

[17] T. P. Van, T. M. Nguyen, N. N. Tran, H. V. Nguyen, L. B. Doan, H. Q. Dao, T. T. Minh, Interpreting the latent space of generative adversarial networks using supervised learning, in: 2020 International Conference on Advanced Computing and Applications (ACOMP), IEEE, 2020. URL: https://doi.org/10.1109%2Facomp50827.2020.00015. doi:`10.1109/acomp50827.2020.00015`.

[18] G. Alain, Y. Bengio, Understanding intermediate layers using linear classifier probes, 2018. `arXiv:1610.01644`.

[19] A. Voynov, A. Babenko, Unsupervised discovery of interpretable directions in the gan latent space, in: Proceedings of the 37th International Conference on Machine Learning, ICML'20, JMLR.org, 2020.