Formal Semantic Controls over Language Models

Danilo S. Carvalho[1,2], **Yingji Zhang**[1]†, **Andre Freitas**[1,2,3]
Department of Computer Science[1], National Biomarker Centre, CRUK-MI[2] - University of Manchester
Idiap Research Institute[3]
United Kingdom, Switzerland
<firstname.lastname>@[postgrad.†]manchester.ac.uk

Abstract

Text embeddings provide a concise representation of the semantics of sentences and larger spans of text, rather than individual words, capturing a wide range of linguistic features. They have found increasing application to a variety of NLP tasks, including machine translation and natural language inference. While most recent breakthroughs in task performance are being achieved by large scale distributional models, there is a growing disconnection between their knowledge representation and traditional semantics, which hinders efforts to capture such knowledge in human interpretable form or explain model inference behaviour. In this tutorial, we examine from basics to the cutting edge research on the analysis and control of text representations, aiming to shorten the gap between deep latent semantics and formal symbolics. This includes the considerations on knowledge formalisation, the linguistic information that can be extracted and measured from distributional models, and intervention techniques that enable explainable reasoning and controllable text generation, covering methods from pooling to LLM-based.

1. Introduction

Despite the recent language models' increasing feats of state-of-the-art performance in a large variety of NLP tasks, there is a growing disconnection between their knowledge representation and traditional semantics, which hinders efforts to capture such knowledge in human interpretable form or explain model inference behaviour. To address this disconnection, numerous approaches have been proposed to approximate deep latent representations to symbolic models grounded on formal linguistics and well-defined mathematical properties. Those approaches are mostly developed over sentence and paragraph models, not only due to computational capacity and cost considerations, but also due to their semantic and structural independence as linguistic units (Allerton, 1969), allowing the representation of relationships between words. Such relationships are a necessary element to improve performance on certain tasks, such as information retrieval and machine translation. Thus targeting them strikes a balance between performance scaling and traceability of the captured knowledge.

Research on this topic has steadily advanced together with the general text embedding efforts (Pragst et al., 2020; Liao, 2021), but has gained increased attention in recent years, due to interpretability, control and safety limitations of state-of-the art, very large language models (LLMs). Thus, a key research question is how to harmonise the flexibility and task delivery provided by large distributional models to the ability to trace its knowledge and behaviour in terms of well-defined formal properties. Sentence and paragraph representation models allow experimentation with a focused scope, bringing a diverse set of contributions with fast turnaround. Some of those contributions are then applied to the larger models (Li et al., 2020), which leads to a positive cycle of improvement. Furthermore, solutions involving explainability and safeguarding of conversational models inevitably touch the matter of compositionality in natural language, which is an important aspect of text representation research.

However, the diversity of contributions in this subject also brings fragmentation of the community awareness to common issues, which causes considerable replication of efforts, terminology inconsistencies and overall missed oportunities. An important step to alleviate such issues would be compiling and structuring the main advances and knowledge gained within this subject, and present them in a summarised form to a broad NLP / distributional semantics public.

With this tutorial, we propose to introduce the field of neuro-symbolic methods in text representation to a broader NLP audience and to promote constructive discussion among researchers in this topic. This will be achieved by presenting an overview of the evolution on symbolic-aware latent representations, focused on sentence embeddings, starting from their pure distributional origins as an extension of word embedding methods (Kiros et al., 2015) and covering their evolving approaches, including tensor pooling, contrastive learning and autoencoders, up to the most recent incorporation of LLMs. We give special attention to the issues of explainability and control, which are of crescent relevance to the NLP community as a whole.

LREC-COLING 2024 Tutorials, pages 50–55

2. Target Audience

This tutorial is targeted at both academics and practitioners who would like to have a better understanding of the interface between formal linguistics and how they manifest within transformer-based models, and the opportunities and challenges brought by extracting and manipulating symbolic properties in latent spaces. The topics are to be presented in a concise and informative way, not diving into minute technical details of the discussed approaches.

Attendees should have a basic understanding of text embeddings, the transformer architecture and a firm undertanding of basic NLP/CL terminology, such as syntax, semantics, part-of-speech and semantic role labeling. A basic understanding of the mathematical foundation on different loss/objective functions and set theory will certainly improve the tutorial experience, but are not required.

3. Outline

The tutorial is organised to follow a *conceptual* and *chronological* order, prioritising the understanding of concepts and then their application. It is divided in the following chapters:

The evolutionary arch from word embeddings to LLMs vs. formal linguistics

We present the motivation and intuition behind the construction of sentence/paragraph embedding models. Starting from their first popularisation as an extension of word embedding models (Kiros et al., 2015) and their applications to the employment of transformer-based architectures (Reimers and Gurevych, 2019; Sanh et al., 2019; Wang et al., 2020; Ni et al., 2022). We explore the characteristics, improvements and shortcomings of the main approaches, contrasting the evolution of distributional semantics with the staticity of formal linguistics, along with the relevant datasets, metrics and benchmarks. This chapter provides a foundation for understanding the topic.

Contrastive learning and conceptual modeling

Considering the most basic goal of obtaining a sentence representation that can be compared to others for measuring semantic similarity, i.e., whether two sentences have similar meaning, it is not suprising that contrastive learning is among the most popular approaches for this end (Tan et al., 2022a; Cheng et al., 2023; Wang et al., 2022; Wu et al., 2022). Contrastive learning works by presenting a set of similar (positive) and dissimilar (negative) examples w.r.t. to a given sample, so that the model learns to place similar ones closer to each other and push apart the dissimilar ones in its latent space.

Another relevant way of learning sentence representations is by leveraging structured knowledge bases of declarative sentences such as definitions, e.g., dictionaries. The intuition in this case being that similar concepts are defined with similar sentences (Hill et al., 2016; Tsukagoshi et al., 2021). Studies on this problem led to the formulation of a NLP task called *definition modeling* dedicated to learning embeddings from definition sentences (Noraset et al., 2017).

This chapter explores the major concepts and relevant works on contrastive learning for sentence representation and conceptual modeling, covering their main achievements and how they are used currently.

Interpretability and formal linguistics

Explainable and interpretable representations are the ones that can be decomposed into factors that are traceable to human understandable concepts. For example, a sentence representation consisting in only two features: the length of the sentence (number of words) and if the sentence is a question or not, is an interpretable one, as both features are easily understood by humans.

Distributed latent embeddings are typically *not interpretable*, which means that inference results obtained from their application are obscure to humans. This limits their application possibilities and brings safety / bias concerns. For this reason, significant attention is being directed towards the creation of explainable representations, specially regarding models dedicated to sensitive tasks or facing the public. Formal syntactic and semantic concepts, such as subject/object and agent/action, provide a strong grounding for the interpretation of latent features if they can be represented in such models.

This chapter deals with different interpretability concerns and approaches, covering the three levels of transparency in explainable AI: algorithmic transparency, decomposability and simulatability, from a text embedding perspective.

Disentanglement and separability

One of the ways to improve explainability is by disentanglement or separation of representations. Disentanglement consists in the separation of traceable factors by binding them to different dimensions (or set of). For example, having the number (singular/plural) of a subject, or time (past/present/future) of a verb strongly tied to a single or limited set of dimensions of the representation. Separability refers to spatially distinguishable clusters in the latent space. For example, having all sentences

with “television” as subject being in a enclosed region in the latent space. Having had significant success in the Computer Vision field, different disentanglement and separability approaches are recently being explored in NLP, notably in sentence representation models (Hu et al., 2017; Chen et al., 2019; Mercatali and Freitas, 2021; Carvalho et al., 2022b).

This chapter explores important concepts regarding the disentanglement/separability of sentence embeddings and how they help achieving explainability.

Control mechanisms for text generation and inference over latent spaces

Most of the current breakthroughs in NLP are related to generative language models, which brought unprecedent levels of attention to such methods both within the NLP community and by the general public. The speed in which this technology has been adopted in a variety of real-world scenarios, from computer programming to medicine, also helped to raise concerns regarding safety, social biases and explainability of the text generated by these systems. Those concerns ultimately translate in the necessity of better control mechanisms over generative models, which are discussed in this chapter, specifically for the case of sentence generation with emphasis on intervention routes through the models' latent spaces, including disentanglement of generative factors (Hu et al., 2017; Mercatali and Freitas, 2021) and linguistic-aware loss functions (Chen et al., 2019).

The role of compositionality in improving representations

One key aspect of condensing sentence information is capturing the relationships between words and how their combination brings forth new meaning: the compositional aspect of language. Compositionality has a pivotal role in the improvement of text representations as the ability to deconstruct relationships such as ellipsis (Wijnholds and Sadrzadeh, 2019) and adjectival modifiers (Carvalho et al., 2022a) can be used to express them in terms of latent space transformations, which provide a mean of linguistic grounded explainability and control.

This chapter discusses central concepts on compositionality, as well as the findings of seminal and recent studies on this subject and their implications.

Employing Autoencoders for efficiency and control

In recent years, Autoencoder architectures became the foundation of a cascade of important contributions to text representation research. They enable the combination of pre-trained encoder and decoder models to learn highly optimised text embeddings (Li et al., 2020), without the need of re-training complex encoders/decoders. Such optimised embeddings can then be analysed and interventions can be applied directly to the Autoencoder latent space (Carvalho et al., 2022b).

In this chapter we explore the benefits and limitations of Autoencoder architectures for sentence embedding and some of their recent developments.

Controlling the semantic properties of large language models

Following the Autoencoder (AE) based developments, we get to the latest incorporation of large language models (LLMs), such as the GPT or LLaMa families, to sentence embedding techniques. While there are still many open research questions regarding the nature of the knowledge embedded in LLM latent spaces, there is a growing consensus on that filtering such knowledge is crucial in enabling their effective and safe use (Meng et al. (2022); Wu et al. (2024); Petroni et al. (2019); Dai et al. (2022), among others), and that it is a certain way of obtaining better text representations (Wijesiriwardene et al., 2024; Zhang et al., 2024). This chapter discusses the main current approaches to achieve semantic control over LLM models, with an emphasis on AE-based studies, but also covering other methods.

Probing sentence latent spaces: geometrical and linguistic properties

Finally, the last chapter discusses techniques for analysis and control of the sentence representations, in particular through intervention to the modeled latent spaces. Namely, different probing methods, and the analysis of geometrical and linguistic properties of the embedding space, such as vector arithmetic, semantic continuity, syntactic and semantic role representation and compositionality. The knowledge gained from all the previous chapters is visited here, so the participants can appreciate the development context of the discussed techniques, as well as their strong and weak points.

Hands-on: Probing Large Language VAEs with LangSpace & LangVAE

In tandem with the discussions on latent space control mechanisms and probing techniques, we demonstrate the applicability and impact of said techniques to current language models hosted in HuggingFace, in a hands-on coding session using our recently developed toolkit. This covers the quick creation and fine tuning of large language

VAEs from stock LLMs, and the probing of created models on predefined tasks using the *LangVAE*[1] and *LangSpace*[2] libraries, respectively.

4. Reading List

Relevant materials to read prior to attending the tutorial include:

- The 2013 review paper: *Representation Learning: A Review and New Perspectives* (Bengio et al., 2013).
- The book *Natural language processing with transformers* (Tunstall et al., 2022)
- The 2020 paper: *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI* (Arrieta et al., 2020).

Further information in the topic can be found in the cited literature and also:

- The book *Representation learning for natural language processing* (Liu et al., 2023).
- Other relevant papers: (Conneau et al., 2018; Kelly et al., 2020; Zhu and de Melo, 2020; Tan et al., 2022b; Opitz and Frank, 2022)

5. Resources

The tutorial resources (slides, code, etc.) will be made available at the web address: https://danilosc.com/events/tutorial-lrec-2024 and by the ACL anthology portal.

6. Presenters

Danilo S. Carvalho is a Principal Clinical Informatician (Research Associate) at the National Biomarker Centre, Cancer Research UK - Manchester Institute, at the University of Manchester, working on Safe and Explainable Artificial Intelligence (AI) architectures. He has experience in both industry and academia, having presented works at multiple international conferences over the past 10 years, such as EACL and ESANN. His main area of expertise is representation learning for NLP and his research interests include explainable AI and legal and patent text processing.

Yingji Zhang is a 3rd year PhD student at the University of Manchester. His research interests include natural language inference, controllable natural language generation, and disentangled representation learning.

Andre Freitas is a Senior Lecturer at the Department of Computer Science at the University of Manchester. He leads the Neuro-symbolic AI group at Idiap and at the Department of Computer Science at the University of Manchester. His main research interests are on enabling the development of AI methods to support abstract, explainable and flexible inference. In particular, he investigates how the combination of neural and symbolic data representation paradigms can deliver better inference. Some of his research topics include: explanation generation, natural language inference, explainable question answering, knowledge graphs and open information extraction.

7. Ethics Statement

The analysis and control of text generation models facing end users need to deal with ethics issues regarding biased and potentially unsafe (offensive, incorrect or misleading) outputs. The tutorial also seeks to inform the participants of these issues and the importance of mitigating them with or without the materials discussed.

References

D.J. Allerton. 1969. The sentence as a linguistic unit. *Lingua*, 22:27–46.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Danilo S Carvalho, Edoardo Manino, Julia Rozanova, Lucas Cordeiro, and André Freitas. 2022a. Montague semantics and modifier consistency measurement in neural language models. *arXiv preprint arXiv:2212.04310*.

[1] https://github.com/neuro-symbolic-ai/LangVAE

[2] https://github.com/neuro-symbolic-ai/LangSpace

Danilo S. Carvalho, Giangiacomo Mercatali, Yingji Zhang, and André Freitas. 2022b. Learning disentangled representations for natural language definitions. In *Findings*.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. A multi-task approach for disentangling syntax and semantics in sentence representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2453–2464.

Qinyuan Cheng, Xiaogui Yang, Tianxiang Sun, Linyang Li, and Xipeng Qiu. 2023. Improving contrastive learning of sentence embeddings from ai feedback. In *Annual Meeting of the Association for Computational Linguistics*.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Annual Meeting of the Association for Computational Linguistics*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR.

Matthew A. Kelly, Yang Xu, Jesús Calvillo, and D. Reitter. 2020. Which sentence embeddings and which layers encode syntactic structure? In *Annual Meeting of the Cognitive Science Society*.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems*, 28.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4678–4699.

Danqi Liao. 2021. Sentence embeddings using supervised contrastive learning. *arXiv preprint arXiv:2106.04791*.

Zhiyuan Liu, Yankai Lin, and Maosong Sun. 2023. *Representation learning for natural language processing*. Springer Nature.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Giangiacomo Mercatali and André Freitas. 2021. Disentangling generative factors in natural language with discrete variational autoencoders. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3547–3556.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Juri Opitz and Anette Frank. 2022. Sbert studies meaning representations: Decomposing sentence embeddings into explainable semantic features. In *AACL*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Louisa Pragst, Wolfgang Minker, and Stefan Ultes. 2020. Comparative study of sentence embeddings for contextual paraphrasing. In *International Conference on Language Resources and Evaluation*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*, pages 3982–3992.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Haochen Tan, Wei Shao, Han Wu, Ke Yang, and Linqi Song. 2022a. A sentence is worth 128 pseudo tokens: A semantic-aware contrastive learning framework for sentence embeddings. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 246–256.

Haochen Tan, Wei Shao, Han Wu, Ke Yang, and Linqi Song. 2022b. A sentence is worth 128 pseudo tokens: A semantic-aware contrastive learning framework for sentence embeddings. *ArXiv*, abs/2203.05877.

Hayato Tsukagoshi, Ryohei Sasano, and Koichi Takeda. 2021. Defsent: Sentence embeddings using definition sentences. In *Annual Meeting of the Association for Computational Linguistics*.

Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. 2022. *Natural language processing with transformers*. " O'Reilly Media, Inc.".

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Yau-Shian Wang, Ashley Wu, and Graham Neubig. 2022. English contrastive learning can learn universal cross-lingual sentence embeddings. *ArXiv*, abs/2211.06127.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Aishwarya Naresh Reganti, Vinija Jain, Aman Chadha, Amit Sheth, and Amitava Das. 2024. On the relationship between sentence analogy identification and sentence structure encoding in large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 451–457, St. Julian's, Malta. Association for Computational Linguistics.

Gijs Jasper Wijnholds and Mehrnoosh Sadrzadeh. 2019. Evaluating composition models for verb phrase elliptical sentence embeddings. In *North American Chapter of the Association for Computational Linguistics*.

Xing Wu, Chaochen Gao, Zijia Lin, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022. Infocse: Information-aggregated contrastive learning of sentence embeddings. In *Conference on Empirical Methods in Natural Language Processing*.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. 2024. Interpretability at scale: Identifying causal mechanisms in alpaca. *Advances in Neural Information Processing Systems*, 36.

Yingji Zhang, Danilo Carvalho, Marco Valentino, Ian Pratt-Hartmann, and Andre Freitas. 2024. Improving semantic control in discrete latent spaces with transformer quantized variational autoencoders. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1434–1450, St. Julian's, Malta. Association for Computational Linguistics.

Xunjie Zhu and Gerard de Melo. 2020. Sentence analogies: Linguistic regularities in sentence embeddings. In *International Conference on Computational Linguistics*.