# Deep Learning for Understanding of Narratives: What Is Still Missing?

Marie-Francine Moens[1]

[1]Department of Computer Science, Celestijnenlaan 200A, 3020 Heverlee, Belgium

**Abstract**

We identify several challenges faced by current deep learning models that aim to understand natural language, particularly when the machine needs to comprehend lengthy discourse such as narratives and in tasks requiring reasoning with content. Many of these issues are equally significant in the processing of visual content and multimodal tasks, such as video understanding or story-to-video generation. To address these challenges, we compare how machines process content with what is known about the language and multimodal processing mechanisms of the human brain. This comparison leads to potential solutions for improving machine understanding, including modifications to model architectures, training methods and inference strategies. These modifications emphasize representations that model contextuality, identity and situational knowledge.

**Keywords**

Representation learning, Situation modeling, Machine understanding of narratives, Story-to-video generation

## 1. Introduction

Understanding is often defined as comprehension, that is, to have a clear or complete idea of the content that is understood.[1] In this paper we aim to reflect on representation learning with deep neural networks for the purpose of natural language understanding (NLU) and beyond. We focus on natural language understanding of narratives. NLU has evolved significantly over time, and it is essential to reflect on its progress both in the past and in the present. However, this progress brings with it challenges, prompting the question: what are the persistent problems in NLU? By examining how humans understand language and the surrounding world, we can gain valuable insights that can guide future advancements in NLU of narratives.[2]

## 2. NLU now

Currently, foundation models serve as a basis for artificial intelligence research. They are trained on extensive datasets, such as natural language data or language paired with images [1]. They are trained in a self-supervised manner by masking and reconstructing the masked content as output (see [2] for an overview). More specifically, they are self-trained with neural models by predicting randomly masked content or by anticipating content in a sequential stream of tokens [1]. Foundation models store in their parameters the patterns they have observed in the training data. Their common transformer architecture enables the incorporation of contextual information refined through its attention weights. The models provide priors on how tokens are combined in a language, or on how visual patches co-occur in the physical world [3]. The resulting content representations contain structural knowledge (e.g., how tokens are syntactically combined, how objects are spatially configured, how actions likely follow each other and what their consequences in time are). Therefore, they hold highly valuable knowledge.

*8th International Workshop on Computational Models of Narrative (CMN'25, Genève)*

sien.moens@kuleuven.be (M. Moens)

http://people.cs.kuleuven.be/~sien.moens/ (M. Moens)

0000-0002-3732-9323 (M. Moens)



[1]Merriam-Webster: to comprehend: to grasp the nature, significance, or meaning of; to contain or hold within a total scope, significance, or amount; to include by construction or implication.

[2]This work is partially based on a keynote lecture given at the Northern Lights Deep Learning Conference 2025.

Deep learning provides continuous vector representations of words and sentences by keeping track of contextual information from large textual corpora. Consequently, the foundation models store the massive content in the latent representations of the network efficiently as vectors with continuous values. Commonly, they are trained with neural transformer architectures characterized by attention layers [4] that model correlations between content elements (e.g., between the sub-words or tokens of a natural language text, between the tokens of language and regions in an image). This processing results in contextual representations of the content elements that effectively capture the meaning of elements where their context acts as constraints. This makes foundation models general knowledge stores of the physical and social world they have observed during training or of domain-specific knowledge when trained on domain-specific sources. The resulting representations form the foundation of many successful tasks such as language generation, and image and video diffusion.

The resulting content representations model contextuality in an excellent way. Note that contextuality has many different meanings, but we adopt the meaning used in linguistics, that is, related to the context of something that provides resources for its appropriate interpretation. Regarding language modeling, the vector representations support the hypothesis of distributional semantics stating that semantically similar words have similar contextual distributions. As a result, foundation models already demonstrate a good degree of conceptualization and generalization through inductive training on large data sets [5]. Even if certain content is not seen in the training data, we can infer its meaning based on other content of which the context is similarly distributed [6, 7]. Apart from syntactic and semantic patterns of language, foundation models store a substantial amount of common knowledge of things in the world, among which are properties, relationships, and interactions of objects, which humans are expected to know, and world knowledge (referring to facts about the people, events and objects seen during training) [8, 9].

The content representations (e.g., the vector representations in the neural network layers obtained after the self-training) form the foundation of many downstream generation and understanding tasks in natural language processing. These tasks require varying levels of model supervision through examples annotated with ground truth labels or by querying the foundation model by instructional prompts. Because of the computational complexity of finetuning a full-fledged foundation model with annotated examples, techniques of parameter efficient finetuning were developed. They focus on the most relevant parameters of the model in the optimization process either by introducing new parameters or freezing the entire model except for selected parts. For example, adapters add extra layers to the existing transformer architecture and only finetune these. An adapter layer uses a bottleneck architecture that projects the input to a smaller dimension, applies a non-linear activation function, and then scales it back up to the input's original dimension [10]. Low-Rank Adaptation (LoRA) injects trainable low-rank matrices into each transformer layer [11]. Recently, we have introduced feature routing functions that aid feature selection through the low-rank bottleneck [12]. Apart from transformer architectures, we currently see a resurgence of recurrent neural networks in the form linear-time sequence modeling (MAMBA) [13]. They integrate the idea of a selective state space model to effectively select data (i.e., focus on or ignore particular inputs) and of a linear RNN to model contextuality and to boost efficiency by a parallel scan algorithm for computing the recurrences.

## 3. NLU then

Natural language understanding (NLU) by a machine is one of the oldest goals of artificial intelligence (AI). Since its early days the field was concerned with building effective representations (e.g., [14]). Past research in natural language understanding has promoted symbolic representations composed of discrete predicates and arguments, often using first-order logic or lambda expressions with variable binding [15]. This formalism can both generalize concepts and relationships, and represent instances of persons, objects and attributes. However, it struggles with scalability due to reliance on handcrafted rules or annotated language for training models that translate natural language into an expressive formal symbolic language covering all possible scenarios. Additionally, symbolic representations

have difficulties in representing the fine-grained contextual meaning of language content [16]. These problems are only aggregated when processing long discourses such as narratives, hence the current focus on relying on large language models or other foundation models to process language.

## 4. What are then the problems?

Foundation models store in their parameters the patterns they have observed in the training data in a very efficient way. They remember the content on which they are trained which facilitates the retrieval of this content, but they do not reach human intelligence, that is, to correctly interpret all possible new situations and circumstances [17], to capture long tail knowledge [18] and to correctly model situations [19]. These properties are crucial for processing lengthy narrative discourse. The retrieval is made more accurate by extending the models, that is, extending the contexts that are captured (as done by, e.g., Gemini 1.5 Pro models), the amount of train data, and consequently the number of trained parameters of the models. True understanding of natural language requires functional linguistic competence, that is, the cognitive abilities needed to truly interpret language and infer its real-world implications [19]. One important ability is situation modelling, which refers to the dynamic tracking of individual protagonists, objects, locations, and events as a (long) narrative or conversation unfolds over time [20]. A situation model is a mental representation of the people, setting, actions and events described in explicit statements in language or inferred through world knowledge [21].

To model their tracking, referential relations that connect entities (such as people and objects) and events across sentences are important. As seen above, foundation models and deep neural networks in general naturally model context and integrate it in their representations. The current models have problems in modeling identities. Identity refers to the distinguishing character or personality of an individual but can be expanded to animals, objects, etc. The models have trouble with making inferences with individual variables and tracking their situation when the narrative unfolds over time [22], whether they are protagonists, objects, locations or events. This problem has resemblance to the "binding problem" of neural networks described by [23], that is, their inability to dynamically and flexibly bind information that is distributed throughout the network and to acquire a compositional understanding of the world in terms of symbol-like entities. As a result, the models need annotated training examples or carefully designed prompts (e.g., as extra supervision to learn a downstream task or as example prompts in an in-context learning setting). For instance, [24] realize recognition of identities and instantiation of variables through prompting with human-drafted examples. Failing instantiation hinders deductive inference and abductive inference [22, 25], which is needed in the understanding of narratives.

We postulate that to really deal with the problem of situation modeling and instantiations of protagonists and objects so these can be tractable as events unfold over time, we need novel ways of training the neural models and to learn representations while keeping the rich contextual representation power. Moreover, while learning such representations, we need to design ways to efficiently and effectively leverage the factual and commonsense knowledge of current foundation models. Finally, as events and their participants change over time, we need contextual representations that dynamically adapt through inference with the individual situations of the discourse. Existing entity-centric representations focus on person names and speakers in a dialogue setting [26] but do not attempt to model situations. Knowledge graphs possibly extended with a graph attention mechanism used for representing entities, relations and situations miss the fine-grained contextual information expressed in a discourse and form only a summarized representation of the discourse (e.g., [27]).

Time and place are important pillars of situation modelling. Situations dynamically change over time through the actions that are performed by persons and other actors, and involved actors and objects change location. True language understanding should be able to answer questions about their location at each moment in time covered by the discourse, about the time an action is performed, about sizes and distances between objects in a certain context, about the durations of actions in a certain context, about the postconditions of actions and their consequences on the location of actors and objects, etc.

Past research mainly focused on translation of language into structured symbolic representations that express spatial and temporal relationships between extracted objects and events or time expressions, respectively (e.g., [20, 28]). Reasoning was conducted using the symbolic representations to deduce the specifics of a situation that often were not explicitly stated in language (through inference mechanisms in a knowledge graph or by means of an external solver). The lack of sufficient annotated training data hindered the widespread adoption of these methods. More importanly, symbolic representations often lack fine-grain contextual semantics that might be different according to the discourse context. Hence, the current interest in neuro-symbolic methods, where the challenge is in effectively integrating neural and symbolic computation for learning and reasoning from raw data. For example, [29] finetune a foundation model for question answering that jointly learns the spatial rules of a game, but these models are restricted to specific domains due to scalability issues (combinatorial explosion in symbol grounding). Finally, the following works translate logical constraints into differential forms, using these constraints in the loss functions [30, 31].

Nowadays, foundation models serve as a backbone for spatial and temporal language understanding, often finetuned for their use in specific domains. Especially the vision-language foundation models have proven to capture spatial commonsense knowledge that is not always present in large language models (e.g., sizes of objects, distances between objects in a certain context, common locations of objects) [32]. It is expected that the emerging video-language foundation models (e.g., MERLOT Reserve [33]), will even better integrate physical knowledge to aid temporal language processing and spatiotemporal inference (e.g., regarding duration of actions, changes in location of objects, etc.). However, foundation models still exhibit many shortcomings when it comes to reasoning with spatial common sense [34]. For example, the inadequate spatial reasoning of models in scene generation by diffusion models, visual question answering, and in vision and language reasoning highlights the need for better representations of spatiotemporal situations [20]. This is especially important for embodied AI models that can interact with the physical environment, but equally crucial in machine understanding of a narrative discourse. Also, the temporal grounding of events on a timeline remains a major problem [28].

Similar problems are encountered in video understanding and text-to-video generation.

The many characters in a movie are involved in different actions and events. Training on long discourses, especially video, is computationally very expensive because of a combinatorial explosion of sequences of possible actions and events. Also in this context we need advanced models that represent identity and situations in a efficient way, without loosing contextuality. Because of the computational complexity of processing video data, learning representations that integrate identity information and are able to represent dynamic situations are even more challenging than when dealing with textual narratives.

In story-to-video generation where we translate a textual story into a video, we currently rely on diffusion models. However it remains very hard to consistently generate the same appearances for the same character in the story. Current research focuses on how to consistently generate the same dog if the story is about my dog (see: https://www.youtube.com/watch?v=vPPryjwTi9w), how to consistently generate the same objects in a scene, how to consistently generate the same background or detect that another background needs to be generated, and how to control changing objects, attributes, locations, postconditions of actions (e.g., [35, 36]), etc. In other words we investigate how to build neural representations that capture identity and their context to be used in the diffusion process when generating the video.

## 5. What can we learn from understanding of language by humans?

The language network of the human brain operates within a receptive window of a few words [37]. It supports the retrieval of individual word meanings and facilitates combinatorial syntactic and semantic processes within that window. This aspect of the human brain is well captured by current large language models.

The default mode network of the human brain is responsible for connecting information across

sentences and to construct a coherent narrative or dialogue representation [37]. Some brain scientists argue that local chunks of linguistic structure are incrementally processed in increasingly abstract levels of representation guided by anticipated content [38, 39, 40, 41].

With regard to identity, cognitive scientists argue that certain human capabilities are innate and learned during evolution. They regard, for instance, the detection and tracking of persons, faces, and animals, as well as inanimate objects and numbering. Humans and certain mammals use a unique name to distinguish between persons, animals and objects of the same type [42, 43].

For a long time, cognitive scientists have identified various types of event indices, including protagonists, temporality, causality, and spatiality [44].

The human brain continuously predicts or samples from a hierarchy of representations when reading or hearing language, where the anticipated content is often visually rendered. One assumes that the temporal receptive windows of language show sensitivity to varying context lengths. The set of hierarchically organized representations evaluate the current text input and might facilitate belief revision by humans when understanding language [43].

The above capabilities of the human brain are not or only partially realized by current NLU models.

## 6. Potential solutions

In this paper we postulate that translating a narrative discourse or natural language in general into a structured representation which comes close to a symbolic representation, such as a knowledge graph that allows embeddings of entities and reasoning based on the graph topology, is not sufficient to solve the above problems, because such representations lose important fine-grained contextual information. On the other hand, we see a lot of potential in expanding current deep learning models with graph based characteristics such as modeling identity, coreference relationships and spatio-temporal semantics.

A first requirement is that identity and coreference information becomes integrated in the representations.

A first way is to add such information in the prompts that are used to query a discourse. In this respect [45] propose to train a model that detects coreferring entities in a dialogue and that adds an identity number representing the entity cluster to entity mentions in the dialogue. The classifier predicts transition functions that guide the generation of the text and its entity cluster in a sequence-to-sequence generation set-up. This approach could also be applied to a narrative discourse. It basically augments the mention of entities in a text with their entity coreference information, after which the representation of the augmented text can be used in a question-answering task.

Another approach is to represent the entities as additional tokens when processing a discourse. This could be realized by modeling a separate memory or by integrating this information as specific learnable tokens or embeddings in the model.

An example of the former is [46]. Here memory storage is explicitly segregated from the computation of the neural network as suggested by [47]. The idea is to store or write long-term and short-term context into a memory. The model iteratively (e.g., with multiple hops) reads relevant information from the memory to answer a question about the discourse. Memory addressing is guided by an attention mechanism usually performed by a controller. Such a method combines successful machine learning strategies for inference with a memory component that can be read and written to. The model is trained to learn how to operate effectively with the memory component. In [46] a model is fed with a stream of data $X$, incrementally processed to fill a memory $M$. Such memory is used to obtain question-related $Q$ clues to provide an answer $A$. Pretext self-supervised tasks improve memorization by continually rehearsing and anticipating coreference information. The memory module consists of a self-attention and a cross-attention layer that fuses the information from input to memory at each time step. It also uses a gating mechanism to allow it to forget. The output decoder takes the question $Q$ and the last memory state $M_T$ as inputs to obtain the answer to a question.

Another inspiring example is found in [48] where the authors model entities and their relations in a text-graph transformer model. They do so by adding new tokens in the textual input, which are

designed to aggregate information via explicit graph relations in the computation of attention weights forming a joint text-graph transformer model.

As the narrative discourse unfolds we need to represent and track entities through time. The field of computer vision might give some inspiration. For instance, object-centric representations of complex scenes in computer vision are instructive [49]. In this work, slot-attention mechanisms provide a differentiable interface between observed entity mention representations and a set of variables called slots. But the approach is currently only applied in simple synthetic settings. A challenging part is to accurately and efficiently model the attributes in the actor- and object-centric representations, their actions and circumstances, that is, modeling their contextual information, so that they minimize redundant information. Another problem asking for solutions regards dealing with entities or objects that appear and disappear in the discourse. When you eat a peeled banana, the banana is gone afterwards. Here recent work that emulates the genesis and evolution of biological cells with an artificial neural network might be inspiring (e.g., [50]).

It remains very challenging to represent dynamic situations of the narrative through time. As the narrative unfolds, some context changes, while at other moments in time the context remains the same. A significant challenge here again is the implicitness of language making the specific location of a person/object at a point in the discourse unobserved. We humans rely on our world knowledge to reconstruct the situation. There is the potential of hidden physics neural networks [51] that work with variables in loss functions that are unobserved or difficult to measure in the training data, so the value of the unobserved variable is inferred from observed data and past states, and the former is then used as ground truth. We might simulate trajectory paths of objects through time with a neural network with the initial state (location) of an object as input, while only a final location is known [52]. During training different paths of possible locations through time could be generated, and the model learns to select one or more trajectories upon the availability of new evidence (in a kind of belief revision way).

The above also entails that the representations that the machine builds of the situation at hand is continually changing as the discourse unfolds, in a way similar to the updates that humans make with regard to their mental representation when comprehending narratives [53]. In the latter case, at significant coherence breaks of the narrative, the representation is fully reconstructed, while during model updating, the existing representation is partially updated with only specific elements adjusted through inference processes. During model construction at major coherence breaks, an event model is entirely reconstructed, whereas during model updating, the existing event model is retained, with only certain elements modified through inference processes.

When the machine processes narrative content, there is also the need to represent and to anticipate the content at different layers of detail and abstraction. In this respect recent unsupervised models that learn to plan for language modeling are inspiring (e.g., [54]) as they can explicitly address the hierarchical nature of language. Although such an approach is developed for text generation, the unsupervised abstraction of content that these models propose is valuable when storing content previously perceived in a discourse or to anticipate what could come next. Typically, such models represent abstract content by embeddings stored in codebooks that might be trained a priori and possibly adapted during training of the language model. However, it remains unclear which content segments are best abstracted and at which level of detail or compression, while keeping this abstraction scalable especially when dealing with a lengthy discourse, and how to dynamically update such representations as the discourse progresses.

The above especially relates to machine understanding of narratives that focuses on grounding of language in the physical world and on comprehending the physical implications of what is communicated in language. Grounding of language in the social world and comprehending its social implications (e.g., [55]) opens another large venue of future research, which we did not cover in this paper.

The proposed solutions could also lead to novel ways of training foundation models that integrate situational knowledge much better than is done now, among which are the learning of representations of individual characters and objects, abstractions of events, and their temporal, spatial and causal relationships.

## 7. Conclusion

We reviewed recent deep learning work on narrative understanding, discussed its shortcomings in handling key narrative phenomena (e.g., object trajectories, identity, and appearance/disappearance), and suggested potential solutions inspired by other domains. The paper gives a good sense of the issues involved with bridging the gap between (a) specialized architectures that are able to handle narrative phenomena well but only in restricted domains or on synthetic text and (b) more general architectures and foundation models that perform well overall but have deficiencies regarding narrative phenomena. We hope that this paper inspires future research on comprehension of narrative content by a machine.

We especially focused on the comprehension of narrative text and the learning of suitable representations for this task. Many of the insights might also be applicable to other tasks such as the machine understanding of narrative video. If we find suitable solutions, the resulting representations could also have an impact on generation tasks such as story-to-video generation.

## Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

## References


[1] R. Bommasani, D. Hudson, E. Adeli, On the opportunities and risks of foundation models, 2022. ArXiv:2108.07258.

[2] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, J. Gao, Large language models: A survey, 2024. ArXiv:2402.06196.

[3] S. Yang, O. Nachum, Y. Du, J. Wei, P. Abbeel, D. Schuurmans, Foundation models for decision making: Problems, methods, and opportunities, 2023. ArXiv:2303.04129.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, NeurIPS, 2017, p. 30.

[5] L. Yang, S. Zhang, L. Qin, Y. Li, Y. Wang, H. Liu, J. Wang, X. Xie, Y. Zhang, Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective, in: Findings of the Association for Computational Linguistics, ACL, 2023, p. 12731–12750.

[6] K. Deschacht, M.-F. Moens, Semi-supervised semantic role labeling using the latent words language model, in: P. Koehn, R. Mihalcea (Eds.), Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, ACL, Singapore, 2009.

[7] K. Deschacht, J. De Belder, M.-F. Moens, The latent words language model, Comput. Speech Lang. 26 (2012) 384–409.

[8] X. Zhou, Y. Zhang, L. Cui, D. Huang, Evaluating commonsense in pre-trained language models, 2020, p. 9733–9740.

[9] E. Davis, Benchmarks for automated commonsense reasoning: A survey, ACM Computing Surveys 56 (2023) 1–41.

[10] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for NLP, in: Proceedings of the 36th International Conference on Machine Learning (ICML 2019) - Proceedings of Machine Learning Research, volume 97, 2019, p. 2790–2799.

[11] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: Proceedings of the International Conference on Learning Representations, 2022.

[12] T. Qu, T. Tuytelaars, M.-F. Moens, Introducing routing functions to vision-language parameter-

efficient fine-tuning with low-rank bottlenecks, in: Proceedings of the 18th European Conference on Computer Vision, ECCV, 2024.
[13] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, 2024. ArXiv: 2312.00752.
[14] R. Schank, R. Abelson, Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures, 1977.
[15] P. Blackburn, J. Bos, Representation and Inference for Natural Language: A First Course in Computational Semantics (Studies in Computational Linguistics), The University of Chicago Press, 2005.
[16] R. Cartuyvels, G. Spinks, M.-F. Moens, Discrete and continuous representations and processing in deep learning: Looking forward, AI Open 2 (2021) 143–159.
[17] W. Nuyts, R. Cartuyvels, M.-F. Moens, Explicitly representing syntax improves sentence-to-layout prediction of unexpected situations, Transactions of the Association for Computational Linguistics 12 (2024) 264–282.
[18] N. Kandpal, H. Deng, A. Roberts, E. Wallace, C. Raffel, Large language models struggle to learn long-tail knowledge, in: Proceedings of the 40th International Conference on Machine Learning, ICML, volume 202. JMLR.org, Article 641, 2023, p. 15696–15707.
[19] K. Mahowald, A. Ivanova, I. Blank, N. Kanwisher, J. Tenenbaum, E. Fedorenko, Dissociating language and thought in large language models, Trends in Cognitive Sciences 28 (2024) 517–540.
[20] P. Kordjamshidi, J. Pustejovsky, M.-F. Moens, Spatial Language Understanding: Representation, Reasoning, and Grounding (in press), 2025.
[21] A. C. Graesser, M. Singer, T. Trabasso, Constructing inferences during narrative text comprehension, Psychological Review 101(3) (1994) 371.
[22] N. Kim, S. Schuster, Entity tracking in language models, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, ACL, 2023, p. 3835–3855.
[23] K. Greff, S. Steenkiste, J. Schmidhuber, On the binding problem in artificial neural networks, 2020. ArXiv preprint arXiv:2012.05208.
[24] W. Wang, T. Fang, C. Li, H. Shi, W. Ding, B. Xu, Z. Wang, J. Bai, X. Liu, C. Jiayang, C. Chan, Y. Song, Candle: Iterative conceptualization and instantiation distillation from large language models for commonsense reasoning, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics - Volume 1: Long Papers, ACL, 2024, p. 2351–2374.
[25] J. Tenenbaum, C. Kemp, T. Griffiths, N. Goodman, How to grow a mind: Statistics, structure, and abstraction, Science 331 (2011) 1279–1285.
[26] L. Aina, C. Silberer, I.-T. Sorodoc, M. Westera, G. Boleda, What do entity-centric models learn? insights from entity linking in multi-party dialogue, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1, ACL, 2019, p. 3772–3783.
[27] P. Veličković, G. Cucurull, A. Casanova, R. A., P. Liò, Y. Bengio, Graph attention networks, in: Proceedings of the International Conference on Learning Representations, ICLR, 2018.
[28] Q. Ning, B. Zhou, H. Wu, H. Peng, C. Fan, M. Gardner, A meta-framework for spatiotemporal quantity extraction from text, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, ACL, 2022, p. 2736–2749.
[29] D. Cunnington, M. Law, J. Lobo, A. Russo, The role of foundation models in neuro-symbolic learning and reasoning, 2024. ArXiv:2402.01889.
[30] A. Leeuwenberg, M.-F. Moens, Temporal information extraction by predicting relative time-lines, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, ACL, Brussels, Belgium, 2018.
[31] A. Leeuwenberg, M.-F. Moens, Towards extracting absolute event timelines from English clinical reports, IEEE/ACM Transactions on Audio, Speech and Language Processing 28 (2020) 2710–2719.
[32] W. Zhao, J. Chiu, C. Cardie, A. Rush, Abductive commonsense reasoning exploiting mutually exclusive explanations, in: Proceedings of the 61st Annual Meeting of the Association for Computational

Linguistics - Volume 1: Long Papers, ACL, 2023, p. 14883–14896.
[33] R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, S. Mohammadreza, A. Kusupati, J. Hessel, A. Farhadi, Y. Choi, MERLOT reserve: Neural script knowledge through vision and language and sound, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, p. 16375–16387.
[34] A. Cohn, An evaluation of ChatGPT-4's qualitative spatial reasoning capabilities in RCC-8, 2023. CoRR, abs/2309.15577.
[35] M. Li, T. Qu, T. Tuytelaars, M.-F. Moens, Towards more accurate personalized image generation: Addressing overfitting and evaluation bias, 2025. URL: https://arxiv.org/abs/2503.06632. arXiv:2503.06632.
[36] M. Trusca, T. Tuytelaars, M.-F. Moens, DM-Align: Leveraging the power of natural language instructions to make changes to images, Computer Vision and Image Understanding 252 (2025) 104292.
[37] E. Fedorenko, A. Ivanova, T. Regev, The language network as a natural kind within the broader landscape of the human brain, Nature Reviews Neuroscience 25 (2024) 289–312.
[38] L. Kristensen, M. Wallentin, Putting Broca's region into context: fMRI evidence for a role in predictive language processing, in: R. Willems (Ed.), Cognitive Neuroscience of Natural Language Use, Cambridge University Press, 2015, p. 160–181.
[39] M. Christiansen, N. Chater, The now-or-never bottleneck: A fundamental constraint on language, The Behavioral and Brain Sciences 39 (2016) 62.
[40] T. Regev, C. Casto, E. Hosseini, M. Adamek, P. Brunner, E. Fedorenko, Intracranial recordings reveal three distinct neural response patterns in the language network, bioRxiv 2022.12.30.522216 (2022).
[41] C. Caucheteux, A. Gramfort, J. King, Evidence of a predictive coding hierarchy in the human brain listening to speech, Nature Human Behaviour 7 (2023) 430–441.
[42] J. Elman, Learning and development in neural networks: The importance of starting small, Cognition 48 (1993) 71–99.
[43] S. Dehaene, How We Learn: The New Science of Education and the Brain, Pinguin Books, 2020.
[44] R. Zwaan, M. C. Langston, A. C. Graesser, The construction of situation models in narrative comprehension: An event-indexing model, Psychological Science 6 (1995) 292–297.
[45] B. Bohnet, C. Alberti, M. Collins, Coreference resolution through a seq2seq transition-based system, Transactions of the Association for Computational Linguistics 11 (2023) 212–226.
[46] V. Araujo, A. Soto, M.-F. Moens, A memory model for question answering from streaming data supported by rehearsal and anticipation of coreference information, in: Findings of the Association for Computational Linguistics, ACL, 2023, p. 13124–13138.
[47] J. Weston, S. Chopra, B. A, Memory networks, in: Proceedings of the 3rd International Conference on Learning Representations, ICLR, 2015.
[48] A. Coman, C. Theodoropoulos, M.-F. Moens, J. Henderson, Gadepo: Graph-assisted declarative pooling transformers for document-level relation extraction, in: Proceedings of the 3rd Workshop on Knowledge Augmented Methods for NLP, ACL, 2024, p. 1–14.
[49] F. Locatello, T. Weissenborn, D. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, T. Kipf, Object-centric learning with slot attention, in: Advances in Neural Information Processing Systems, NeurIPS, volume 33, 2020, p. 11525–11538.
[50] E. Nisioti, E. Plantec, M. Montero, J. Pedersen, S. Risi, Growing artificial neural networks for control: The role of neuronal diversity, in: Proceedings of the Genetic and Evolutionary Computation Conference Companion, GECCO, ACM, 2024, p. 175–178.
[51] M. Raissi, A. Yazdani, G. Karniadakis, Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations, Science 367 (2020) 1026–1030.
[52] C. Legaard, T. Schranz, G. Schweiger, J. Drgoňa, B. Falay, C. Gomes, A. Iosifidis, M. Abkar, P. Larsen, Constructing neural network based models for simulating dynamical systems, ACM Computing Surveys 55 (2023).
[53] I. R. Brich, F. Papenmeier, M. Huff, M. Merkt, Construction or updating? Event model processes

during visual narrative comprehension, Psychonomic Bulletin and Review 31 (5) (2024).
[54] N. Cornille, M.-F. Moens, F. Mai, Learning to plan for language modeling from unlabeled data, in: Proceedings of the Conference on Language Modeling, COLM, 2024.
[55] L. Allein, M. M. Trusca, M.-F. Moens, Interpretation modeling: Social grounding of sentences by reasoning over their implicit moral judgments, Artificial Intelligence 338 (2025).