Modular Mechanistic Networks: On Bridging Mechanistic and Phenomenological Models with Deep Neural Networks in Natural Language Processing

Simon Dobnik
CLASP and FLOV
University of Gotenburg, Sweden
simon.dobnik@gu.se

John D. Kelleher
ADAPT Centre for Digital Content Technology
Dublin Institute of Technology, Ireland
john.d.kelleher@dit.ie

Abstract

Natural language processing (NLP) can be done using either top-down (theory driven) and bottom-up (data driven) approaches, which we call *mechanistic* and *phenomenological* respectively. The approaches are frequently considered to stand in opposition to each other. Examining some recent approaches in deep learning we argue that deep neural networks incorporate both perspectives and, furthermore, that leveraging this aspect of deep learning may help in solving complex problems within language technology, such as modelling language and perception in the domain of spatial cognition.

1 Introduction

There are two distinct methodologies to build computational models of language or of world in general. The first approach can be characterised as qualitative, symbolic and driven by domain theory (we will call this a *top-down* or *mechanistic approach*), whereas the second approach may be characterised as quantitative, numeric and driven by data and computational learning theory (we will call this the *bottom-up* or *phenomenological approach*). In this context we are borrowing the terminology of *phenomenological model* from the literature on the Philosophy of Science where the term *phenomenological model* is sometimes used to describe models that are independent of theory (see for example (McMullin, 1968)), but more generally is used to describe models that focus on the observable properties (phenomena) of a domain (rather than explaining the hidden mechanisms relating these phenomena) (Frigg and Hartmann, 2017). For this paper we use the term phenomenological model to characterise models which are primarily driven by fitting to observable relationships between phenomena in a domain, as represented by correlations between features in a dataset sampled from the domain; as opposed to models that are derived from a domain theory of the interactions between domain features. The focus of this paper is to examine and frame the potentially synergistic relationship between these distinct analytic methods for natural language processing (NLP) in the light of recent advances in deep neural networks (DNNs) and deep learning.

In historic terms this discussion is recurrent throughout the history of NLP. For example, early approaches such as (Shieber, 1986; Alshawi, 1992) are mechanistic in nature as they are based on logic and other formal approaches such as features structures and unification which are tools that allow formalisation of domain theories. With the availability of large corpora in mid-1990s there was a shift to data-driven phenomenological approaches with a focus on statistical machine learning methods (Manning and Schütze, 1999; Turney et al., 2010). This inspired several discussions on the relation between the two approaches (e.g., (Gazdar, 1996; Jones et al., 2000)). We share the view of some that both approaches are in fact in a complimentary distribution with each other as shown in Table 1 (adapted from a slide by Stephen Pulman). Mechanistic approaches provide deep coverage but of a limited domain; outside a domain they prove brittle and therefore limited. On the other hand, phenomenological approaches are wide-coverage and robust to variation found in data but provide a shallow representation of language.

Our desiderata is a wide-coverage system with deep analyses. It was considered that this could be achieved by a hybrid model but working out such a model has proven not a trivial task. Systems that used both approaches treated them normally as in-

tech/cov	wide	narrow
deep	**our goal**	symbolic
shallow	data-based	**useless**

Table 1: Properties of mechanistic and phenomenological approaches in NLP

dependent black-boxes organised in layers (e.g. (Kruijff et al., 2007)). However, the marked recent advances in the NLP based on *deep* (!) neural networks have made the question of how these two methodologies should be used, related and integrated in NLP research apposite.

The choice of a method depends on the goal of the task for which it is used. One goal for processing natural language is to develop useful applications that help humans in their daily life, for example machine translation and speech recognition. In application scenarios where a rough analysis is acceptable (e.g., a translation that provides the gist of the message) and large annotated and structured corpora are available, machine learning is the methodology of choice to address this goal. However, where precise analysis is required or where there is a scarcity of data, a machine learning approach may not be suitable. Furthermore, if the goal of processing language is rather motivated by the desire to better understand its cognitive foundations, than a machine learning methodology, particularly one based on an unconstrained, fully connected deep neural network, is not appropriate. The criticisms of unconstrained neural network based models (typically characterised by fully-connected feed-forward multi-layer networks) in cognitive science has a long history (see (Massaro, 1988) *inter alia*) and often focuses on (i) the difficultly in analysing in a domain-theoretic sense how the model works, and (ii) the, somewhat ironic, scientific short-coming that neural networks are such powerful and general learning mechanisms that demonstrating the ability of a network to learn a particular mapping or a function is scientifically useless from a cognitive science perspective. In particular, as Massaro (1988) argues, a neural network model is so adaptable that given the appropriate dataset and sufficient time and computing power it is likely to be able to learn mappings that not only support a cognitive theory but also ones that contradict that theory. One approach to address this problem is to introduce domain relevant structural constraints into the model via the network architecture, early approaches include (Feldman et al., 1988; Feldman, 1989; Regier, 1996). Indeed, we argue in this paper that one of the important and somewhat overlooked factors driving the success of research in deep learning is the specificity and modularity of deep learning architectures to the tasks they are applied too.

Contribution: In this paper we evaluate the relation between mechanistic and phenomenological models and argue that although it appears that the former have lost their significance in computational linguistics and its applications they are still very much present in the form of formal language modelling that underlines most of the current work with machine learning. Moreover, we highlight that many of the recent advances in deep learning for NLP are not based on unconstrained neural networks but rather that these networks have task specific architectures that encode domain-theoretic considerations. In this light, the relationship between mechanistic and phenomenological models can be viewed as potentially more synergistic. Given that many logical theories are defined in terms of *functions* and *compositional* operations and neural networks learn and compose functions, a logic-based domain theory of linguistic performance can naturally inform the structural design of deep learning architectures and thereby merge the benefits of both in terms of model interpretability and performance.

Overview: In Section 2, we discuss recent developments in deep learning approaches in NLP and situate them within the current debate; then, in Section 3, we use the computational modelling of spatial language as an NLP case study to frame the possible synergies between formal models and machine learning and set out our thoughts for potential approaches to developing a more synergistic understanding of the formal models and machine learning for NLP research. In Section 4 we give our concluding thoughts.

2 Deep Learning: A New Synthesis?

In recent years deep learning (DL) models have improved or in some cases markedly improved the state of the art across a range of NLP tasks. Some of the drivers of DL success include: (i) the availability of large datasets, (ii) more powerful computers, and (iii) the power of learning and adapt-

ability of connectionist neural networks. However, another and less obvious driver of DL is the fact that (iv) DL network models often have architectures that are specifically tailored or structured to the needs of a specific domain or task. This fact becomes obvious when one considers the variety of DL architectures that have been proposed in the literature. For example, a schematic overview of neural network architectures can be found at at: http://www.asimovinstitute.org/neural-network-zoo/ (van Veen, 2016).

2.1 Modularity in Deep Learning Architectures

There are a large-number of network design parameters that may be driven by experimental results rather than domain theory. For example, (i) the size of the network, (ii) the depth of the layers, (iii) the size of the matrices passed between the layers, (iv) activation functions and (v) optimiser are all network parameters that are often determined through an empirical trial-and-error process that is informed by designer intuition (Jozefowicz et al., 2016). However, the diversity of current network architectures extends beyond differences in these parameters and this diversity of network architecture is not a given. For example, given the flexibility of neural networks, one approach to accommodating structure into the processing of a network is to apply minimal constraints on the architecture and to rely on the ability of the learning algorithm to induce the relevant structure constraints by adjusting the network's weights.

On the other hand, it has, however, long been known that pre-structuring a neural network by the careful design of its architecture to fit the requirements of the task results in better generalisation of the model beyond the training dataset (LeCun, 1989). Understood in this context, DL is assisted (or supervised!) by the task designer in terms of a priori background knowledge who decides what kind of networks they are going to build, the number of layers, what kind of layers, the connectivity between the layers and other parameters. DL is most frequently not using fully connected layers, instead several kinds of layered networks have been developed tailored to the task. In this respect DL models capture top-down domain informed specification that we have seen with the rule-based NLP systems. This flexibility of neural networks is ensured by their modular design which takes as a basis a single perceptron unit which can be thought of encoding a simple concept. When several units are organised and connected into larger collections of units, these may be given interpretations that we give to symbolic representations in rule-based systems. The level of conceptual supervision may thus vary from no-supervision when fully connected layers are used, to weak supervision that primes the networks to learn particular structures, to strong supervision where the structure is given and only parameters of this structure are trained.

An example of weak supervision are Recurrent Neural Networks (RNNs) that capture sequence learning required for language models. The design of current state-of-the-art RNN language models is informed by linguistic phenomena such as short- and long-distance dependencies between linguistic units. In order to improve the ability of RNNs to model long-distance dependencies, contemporary RNN language models use Long-Short Memory Units (LSTM) or Gated Recurrent Units (GRUs) which may be further augmented with attention mechanisms (Salton et al., 2017). The inputs and outputs of such networks can be either characters or words, the latter represented as word embeddings in vector spaces.

Another example of weakly supervised neural networks, in the sense that their design is informed by a domain, are Convolutional Neural Networks (CNNs) which have their origin in image processing (LeCun, 1989). In CNNs the convolutions are meant as filters that encode a region of pixels into a single neural unit which learns to respond to the occurrence of a pixel pattern in the region specific visual feature. Importantly, the weights associated with a specific convolution are shared across a group of neurons such that together the group of neurons check for the occurrence of the visual features across the full surface of the image. Additionally, as objects or entities may occur in different parts of image, to decrease the effects of spatial continuum, operations such as pooling are used that encode convolved representations from various parts of the image. In analogy to learning visual features, CNNs have also been used for language modelling to capture different patterns of characters in strings (Kim et al., 2016).

Specialised networks may be treated as modules which are sequenced after each other. For example, the current Neural Machine Transla-

tion (NMT) architecture is the encoder-decoder (Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015; Kelleher, 2016). This architecture uses one RNN, known as the encoder, to fully process the input sentence and generate its vector based representation. This is passed to a second RNN, the decoder, which implements a language model of the target language which generates the translation word by word. Domain theoretic considerations have affected the design how the two language modelling networks are connected in a number of ways. For example, an understanding that different languages have different word orders lead to enabling the decoder to look both back and forward along the input sentence during translation. This is implemented by fully processing the input sequence with the first RNN before translation is generated by the second RNN. However, the understanding of the need for local dependencies between different sections of the translation and somewhat a contrary requirement to the need for a potentially global perspective on the input has resulted in the development of attention mechanisms within the NMT framework. This means that DL network architectures modules are not only sequenced but they are also stacked. A variant of the NMT encoder-decoder architecture that replaces the encoder RNN with a CNN has revolutionised the field of image captioning (Xu et al., 2015). Figure 1 gives a schematic representation of such image captioning systems. The CNN module learns to represent images as vector representations of visual features and the RNN module is a language model whose output is conditioned on the visual representations. We have already mentioned that CNNs are also used to generate word representations. These representations are then passed to an RNN model to predict the next word in the context of preceding words in the sequence (see (Kim et al., 2016)). The advantage of using a CNN module to learn word representation is that it enables the system to capture spelling variation of morphologically-rich languages or texts from social media that does not use standard spelling of words. This and also the preceding examples therefore illustrate how different levels of linguistic representations are modelled in modular DL architectures.

In summary, the design of a DL architectures, where DL networks are treated as composable modules, can constrain and guide a number of factors that are important in representing language and other modalities, in particular the hierarchical composition of features and the sequencing of the representations. Importantly, the neural representations that are used in these cases are inspired by rich work on top-down rule-based mechanistic natural language processing.

2.2 Phenomenological versus Mechanistic Models

The ability to treat neural networks as composable modules within an overall system architecture is a powerful one. This is because during training it is possible to back-propagate the error through each of the system's modules (networks) and train them in consort while permitting each module to learn its distinctive task in parallel with the other modules in the network. However, the power of this approach has led to some research being based on a relatively shallow understanding of domain theory and most of the work being spent on fitting the hyper-parameters of the training algorithm through a grid-search driven by experimental performance on gold-standard datasets. The domain theory is only used to inform the broad outlines of the system architecture. Using image-captioning as an example, and at the risk of presenting a caricature, this approach may be described as: "we are doing image-captioning so we need a CNN to encode the image and an RNN to generate the language and we will let the learning algorithm sort out the rest of the details".

This theory free, or at least, theory light approach to NLP research is primarily driven by performance on gold-standard datasets and lamentably frequently the analysis of the systems is limited to the presentation of system results relative to a state-of-the-art leader-board with relatively little reflection on the how the structure of the model reflects theoretic considerations. This focus on performance in terms of accurately modelling the empirical relationship between inputs and outputs and where the trained model is treated as a black box aligns with what we describe as the *phenomenological tradition* in machine learning. This can be contrasted with an alternative tradition within machine learning which is sometimes described as being based on *mechanistic models*. Mechanistic models presuppose a domain theory and the model is essentially a computational implementation of this domain theory.

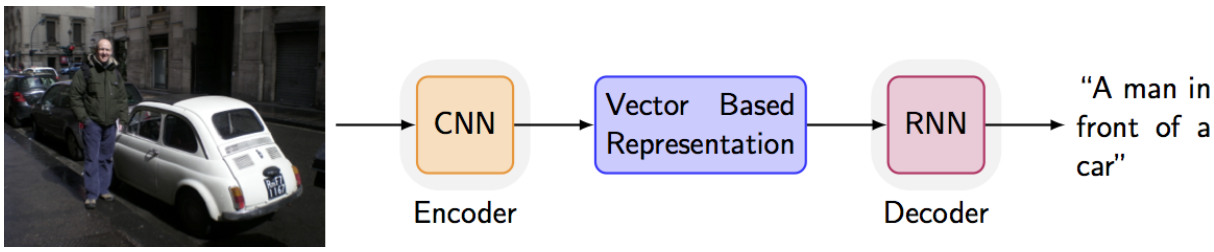

Figure 1: A schematic representation of DL image captioning architectures

To illustrate this difference, contrast for example the approach to training a support vector machine classifier where multiple kernels are tested until one with high performance on a dataset is found versus the approach to defining the topology of a Bayesian network in such a way that it mirrors a theory informed model of the causal relationships between relevant variables in the domain (Kelleher et al., 2015). Once the theoretical model has been implemented, the free parameters of the model can then be empirically fit to the data.

Consequently, mechanistic models are informed by both top-down theoretical considerations of a task designer but they are also sensitive to bottom-up empirical considerations, the training data. Mechanistic models have several advantages, for example: they can be used to test a domain theory. If the model is accurate, this provides evidence that the theory is correct. Assuming the theory is correct, they are likely to outperform phenomenological models in contexts where data is limited.[1] The top top-down approach provides background knowledge that restricts the size of the training search space.

Traditionally, neural networks have been considered the paradigmatic example of a phenomenological model. However, viewing neural networks as component modules within a larger deep-learning systems opens the door to sophisticated mechanistic deep-learning models. Such an approach to network design is, however, dependent on the system designer being informed by domain theory and is therefore strongly supervised in terms of background knowledge. An example of modular networks where each module is some configuration of neural units that are tailored to optimise parameters of a particular task is described in (Andreas et al., 2016) who work in the domain of question answering. The architecture learns how to map questions and visual or database representations to textual answers. In order to answer a question, the network learns a network layout of modules that are responsible for the individual steps required to answer the question. For example, to answer "What colour is the bird" the network applies the attention module to find the object from the question, followed by a module that identifies the colour of the attended region in the image. The possible sequences of modules are constrained by being represented as typed functions: in fact the modules translate to typed functional applications through which compositionality of linguistic meaning is ensured as in formal semantics (Blackburn and Bos, 2005). The system learns (using reinforcement learning) a layout model which predicts the sequence of modules to produce an answer for a question sentence and an execution module which learns how to ground a network layout in the image or database representation. An extension of this work is described in (Johnson et al., 2017) where both procedures rely on less background knowledge. For example, the system does not use a dependency parser to parse the input sentence but an LSTM language module and the modules use a more generic architecture.

The modular networks are in line with the *structured connectionism* of (Feldman et al., 1988) and *constrained connectionism* of Regier "in which complex domain-specific structures are built into the network, constraining its operation in clearly understandable and analysable ways" (Regier, 1996, p. 2). Regier's presentation of constrained connectionism is based on a case study on learning spatial relations and events. The case study describes the design and training of a neural network that receives short movies of 2 two-dimensional objects, a static rectangle and a circle which is either static or moving, as input and the model learns to predict the correct spatial term to describe the position and movement of the circle relative to the

[1] See discussion on generative versus discriminative models in (Kelleher et al., 2015).

rectangle. For example, a static circle might be described as *above* the rectangle, whereas a moving circle might move *out from under* the rectangle. A crucial aspect of this case study for Regier's argument is that the neural network's architecture is constrained in so far as it incorporates a number of structural devices that are motivated by neurological and psychological evidence concerning the human visual system, including motion buffers, angle and orientation computations components, and boundary and feature maps for objects in the input. Following (Regier, 1996), in the next section we will take spatial language as an NLP case-study and discuss how domain theory can be used to extend current deep-learning systems so as to move them further towards the mechanistic pole within the phenomenological versus mechanistic spectrum.

3 Spatial Language

Our focus is computational modelling of spatial language, such as *the chair is to the left and close to the table* or *go down the corridor until the large painting on your right, then turn left*, which requires integration of different sources of knowledge that affect its semantics, including: (i) scene geometry, (ii) perspective and perceptual context, (iii) world knowledge about dynamic kinematic routines of objects, and (iv) interaction between agents through language and dialogue and with the environment through perception. Below we describe these properties in more detail:

Scene geometry is described within a two-dimensional or three-dimensional coordinate frame in which we can represent locations of objects as geometric shapes as well as angles and distances between them. Over a given area we can identify different degrees of applicability of a spatial description, for example with spatial templates (Logan and Sadler, 1996; Dobnik and Åstbom, 2017). A spatial template may be influenced by perceptual context through the presence of other objects in the scene known as distractors (Kelleher and Kruijff, 2005b; Costello and Kelleher, 2006), occlusion (Kelleher and van Genabith, 2006; Kelleher et al., 2011), and attention (Regier and Carlson, 2001).

Directionals such as *to the left of* require a model of *perspective* or *assignment of a frame of reference* (Maillat, 2003) which includes a viewpoint parameter. The viewpoint may be defined linguistically *from your view* or *from there* but it is frequently left out. Ambiguity with respect to the intended perspective of a reference can affect the grounding of spatial terms in surprising ways (Carlson-Radvansky and Logan, 1997; Kelleher and Costello, 2005). However, frequently the intended perspective can be either inferred from the perceptual context (if only one interpretation is possible, see for example the discussion on contrastive versus relative meanings in (Kelleher and Kruijff, 2005a)) or it may be linguistically negotiated and aligned between conversational partners in dialogue (Dobnik et al., 2014, 2015, 2016).

As mentioned earlier, spatial descriptions do not refer to the actual objects in space but to conceptual geometric representations of these objects, which may be points, lines, areas and volumes. The representation depends on how we view the scene, for example *under the water* (water $\approx$ surface) and *in the water* (water $\approx$ volume). The influence of *world knowledge* goes beyond object conceptualisation. Some prepositions are more sensitive to the way the objects interact with each (their dynamic kinematic routines) while other are more sensitive to the way the objects relate geometrically (Coventry et al., 2001).

Finally, because situated agents are located within dynamic linguistic and perceptual environments they must continuously adapt their understanding and representations relative to these context. On the language side they must maintain language coordination with dialogue partners (Clark, 1996; Fernández et al., 2011; Schutte et al., 2017; Dobnik and de Graaf, 2017). A good example of adaptation of contextual meaning through linguistic interaction is the coordinated assignment of frame of reference mentioned earlier.

In summary, the meaning of spatial descriptions is dynamic, dependent on several sources of contextually provided knowledge which provide a challenge for its computational modelling because of its contextual underspecification and because it is difficult to provide and integrate that kind of knowledge. On the other hand, a computational system taking into account these meaning components in context would be able to understand and generate better, more human-like, spatial descriptions and engage in more efficient communication in the domain of situated agents and humans. Furthermore, it could exploit the synergies between different knowledge sources to compensate miss-

ing knowledge in one source from another (Steels and Loetzsch, 2009; Skočaj et al., 2011; Schutte et al., 2017).

3.1 Modular Mechanistic (Neural) Models of Spatial Language

The discussion in the preceding section highlighted the numerous factors that impinge on the semantics of spatial language. It is this multiplicity of factors that make spatial language such a useful case study for this paper, the complexity of the problem invites a modular approach where the solution can be built in a piecewise manner and then integrated. One challenge to this approach to spatial language is the lack of an overarching theory explaining how these different factors should be integrated, examples of candidate theories that could act as a starting point here include (Herskovits, 1987) and (Coventry and Garrod, 2005).

At the same time there are a number of examples of neural models in the literature that could provide a basis for the design of specific modules. We have already discussed (Regier, 1996) which captured geometric factors and paths of motion. Another example of a mechanistic neural model of spatial descriptions is described in (Coventry et al., 2005). Their system processes dynamic visual scenes containing three objects: a teapot pouring water into a cup and the network learns to optimise, for each temporal snapshot of a scene, the appropriateness score of a spatial description obtained in subject experiments. The idea behind these experiments is that descriptions such as *over* and *above* are sensitive to a different degree to geometric and functional properties of a scene, the latter arising from the interactions between objects as mentioned earlier. The model is split into three modules: (i) a vision processing module that deals with detection of objects from image sequences that show the interaction of objects, the tea pot, the water and the cup, using an attention mechanism, (ii) an Elman recurrent network that learns the dynamics of the attended objects in the scene over time, and (iii) a dual feed-forward vision and language network to which representations from the hidden layer of the Elman network are fed and which learns how to predict the appropriateness score of each description for each temporal configuration of objects. Each module of this network is dedicated to a particular task: (i) to recognition of objects, (ii) to follow motion of attended objects in time and (iii) to integration of the attended object locations with language to predict the appropriateness score, factors that have been identified to be relevant for computational modelling of spatial language and cognition through previous experimental work (Coventry et al., 2001). The example shows the effectiveness of representing networks as modules and their possibility of joint training where individual modules constrain each other.

The model could be extended in several ways. For example, contemporary CNNs and RNNs could be used which have become standard in neural modelling of vision and language due to their state-of-the-art performance. Secondly, the approach is trained on a small dataset of artificially generated images of a single interactive configuration of three objects.[2] An open question is how the model scales on a large corpus of image descriptions (Krishna et al., 2017) where considerable noise is added. There will be several objects, their appearance and location may be distorted by the angle at which the image is taken, there are no complete temporal sequences of objects and the corpora typically does not contain human judgement scores on how appropriate a description is given an image. Finally, Coventry et al.'s model integrates three modalities used in spatial cognition, but as we have seen there are several others. An important aspect is grounded linguistic interaction and adaptation between agents. For example, (Lazaridou et al., 2016) describe a system where two networks are trained to perform referential games (dialogue games performed over some visual scene) between two agents. In this context, the agents develop their own language interactively. An open research question is whether parameters such frame of reference intended by the speaker of a description could also be learned this way. Note that this is not always overtly specified, e.g. *from my left*.

Sometimes a mechanistic design of the network architecture constrains what a model can learn in undesirable ways. For example, Kelleher and Dobnik (2017) (in this volume) argue that contemporary image captioning networks as in Figure 1 have been configured in a way that they capture visual properties of objects rather than spatial relations between them. Consequently, within the captions generated by these systems the rela-

[2] To be fair to the authors, their intention was not to build an image captioning system but to show that modular networks can optimise human experimental judgements.

tion between the preposition and the object is not grounded in geometric representation of space but only in the linguistic sequences through the decoder language model where the co-occurrence of particular words in a sequence is estimated. (Dobnik and Kelleher, 2013, 2014) show that a language model is predictive of functional relations between objects that spatial relations are also sensitive to but in this case the geometric dimension is missing. This indicates that the architecture of these image-captioning systems, although modular, ignores important domain theoretic considerations and hence are best understood as close to the phenomenological (black-box) than the mechanistic (grey-box) network design philosophy this paper advocates.

In summary, it follows that an appropriate computational model of spatial language should consist of several connected modalities (for which individual neural network architectures are specified) but also of a general network that connects these modalities, thus akin to the specialised regions and their interconnections in the brain (Roelofs, 2014). The challenge of creating and training such a system is obviously significant, however one feature of neural network training that may make this task easier is that it is possible to back-propagate through a pre-trained network. This opens the possibility of pre-training networks as modules (sometimes even on different datasets) that carry out specific theory-informed tasks and then training larger systems that represent the full-theory by including these pre-trained modules components within the system and training other modules and/or integration layers while keeping the weights of the pre-trained modules frozen during training.

4 Conclusion and Future Research

DNNs provide a platform for machine learning that permits great flexibility in combining top-down specification (in terms of hand-designed structures and rules) and data driven approaches. Designers can tailor the network structures to each individual learning problem and therefore effectively reach the goal of combining mechanistic and phenomenological approaches: a problem that has been investigated in NLP for several decades. The strength of DNNs is in the compositionality of perceptrons or neural units, and indeed networks themselves, which represent individual classification functions that can be combined in novel ways. This was not possible with other approaches in machine learning to the same degree with a consequences that these worked more as black boxes. Finally, although we are not advocating that there is a direct similarity between DNNs and human cognition, it is nonetheless the case that DNNs are inspired by neurons and connectionist organisation of human brain and hence at some high abstract level they share some similarities, for example basic classification units combine to larger structures, the structures get specialised to modules to perform certain tasks, and training and classification is performed across several modules. Therefore, this might be a possible explanation that DNNs have been so successful in computational modelling of language and vision, the surface manifestations of the underlying human cognition, as at some abstract level they represent a similar architecture to human cognition.

Acknowledgements

The research of Dobnik was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at Department of Philosophy, Linguistics and Theory of Science (FLoV), University of Gothenburg.

The research of Kelleher was supported by the ADAPT Research Centre. The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Funds.

References

Hiyan Alshawi. 1992. *The Core Language Engine*. ACL-MIT Press series in natural language processing. MIT Press, Cambridge, Mass.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Learning to compose neural networks for question answering. In *Proceedings of NAACL-HLT 2016*. Association for Computational Linguistics, San Diego, California, pages 1545–1554.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *arXiv preprint*. International Conference on Learning Representations, arXiv:1409.0473v7 [Cs.CL], pages 1–15.

Patrick Blackburn and Johan Bos. 2005. *Representation and inference for natural language. A first course in computational semantics*. CSLI Publications.

L.A. Carlson-Radvansky and G.D. Logan. 1997. The influence of reference frame selection on spatial template construction. *Journal of Memory and Langauge* 37:411–437.

Herbert H. Clark. 1996. *Using language*. Cambridge University Press, Cambridge.

Fintan Costello and John D. Kelleher. 2006. Spatial prepositions in context: The semantics of *Near* in the presense of distractor objects. In *Proceedings of the 3rd ACL-Sigsem Workshop on Prepositions*. pages 1–8.

Kenny Coventry and Simon Garrod. 2005. *Spatial prepositions and the functional geometric framework. Towards a classification of extra-geometric influences.*, volume 2. Oxford University Press.

Kenny R. Coventry, Angelo Cangelosi, Rohanna Rajapakse, Alison Bacon, Stephen Newstead, Dan Joyce, and Lynn V. Richards. 2005. Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In Christian Freksa, Markus Knauff, Bernd Krieg-Brückner, Bernhard Nebel, and Thomas Barkowsky, editors, *Spatial Cognition IV. Reasoning, Action, Interaction*, Springer Berlin Heidelberg, volume 3343 of *Lecture Notes in Computer Science*, pages 98–110.

Kenny R. Coventry, Mercè Prat-Sala, and Lynn Richards. 2001. The interplay between geometry and function in the apprehension of Over, Under, Above and Below. *Journal of Memory and Language* 44(3):376–398.

Simon Dobnik and Amelie Åstbom. 2017. (Perceptual) grounding as interaction. In Volha Petukhova and Ye Tian, editors, *Proceedings of Saardial – Semdial 2017: The 21st Workshop on the Semantics and Pragmatics of Dialogue*. Saarbrücken, Germany, pages 17–26.

Simon Dobnik and Erik de Graaf. 2017. KILLE: a framework for situated agents for learning language through interaction. In Jörg Tiedemann and Nina Tahmasebi, editors, *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)*. Northern European Association for Language Technology (NEALT), Association for Computational Linguistics, Gothenburg, Sweden, pages 162–171.

Simon Dobnik, Christine Howes, Kim Demaret, and John D. Kelleher. 2016. Towards a computational model of frame of reference alignment in Swedish dialogue. In Johanna Björklund and Sara Stymne, editors, *Proceedings of the Sixth Swedish language technology conference (SLTC)*. Umeå University, Umeå, pages 1–3.

Simon Dobnik, Christine Howes, and John D. Kelleher. 2015. Changing perspective: Local alignment of reference frames in dialogue. In Christine Howes and Staffan Larsson, editors, *Proceedings of goDIAL – Semdial 2015: The 19th Workshop on the Semantics and Pragmatics of Dialogue*. Gothenburg, Sweden, pages 24–32.

Simon Dobnik and John Kelleher. 2014. Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes. In *Proceedings of the Third V&L Net Workshop on Vision and Language*. Dublin City University and the Association for Computational Linguistics, Dublin, Ireland, pages 33–37.

Simon Dobnik and John D. Kelleher. 2013. Towards an automatic identification of functional and geometric spatial prepositions. In *Proceedings of PRECogSsci 2013: Production of referring expressions - bridging the gap between cognitive and computational approaches to reference*. Berlin, Germany, pages 1–6.

Simon Dobnik, John D. Kelleher, and Christos Koniaris. 2014. Priming and alignment of frame of reference in situated conversation. In Verena Rieser and Philippe Muller, editors, *Proceedings of DialWatt – Semdial 2014: The 18th Workshop on the Semantics and Pragmatics of Dialogue*. Edinburgh, pages 43–52.

J. A. Feldman, M. A. Fanty, and N. H. Goodard. 1988. Computing with structured neural networks. *Computer* 21(3):91–103.

Jerome A. Feldman. 1989. Structured neural networks in nature and in computer science. In Rolf Eckmiller and Christoph v.d. Malsburg, editors, *Neural Computers*, Springer, Berlin, Heidelberg, pages 17–21.

Raquel Fernández, Staffan Larsson, Robin Cooper, Jonathan Ginzburg, and David Schlangen. 2011. Reciprocal learning via dialogue interaction: Challenges and prospects. In *Proceedings of the IJCAI 2011 Workshop on Agents Learning Interactively from Human Teachers (ALIHT)*. Barcelona, Catalonia, Spain.

Roman Frigg and Stephan Hartmann. 2017. Models in science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy (Spring 2017 Edition)*, Metaphysics Research Lab, Stanford University.

Gerald Gazdar. 1996. Paradigm merger in natural language processing. In Ian Wand and Robin Milner, editors, *Computing Tomorrow*, Cambridge University Press, New York, NY, USA, pages 88–109.

Annette Herskovits. 1987. *Language and Spatial Cognition*. Cambridge University Press, New York, NY, USA.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Fei-Fei Li, C. Lawrence Zitnick, and Ross B. Girshick. 2017. Inferring and

executing programs for visual reasoning. In *arXiv preprint*. arXiv:1705.03633v1 [cs.CV], pages 1–13.

Karen I. B. Spärck Jones, Gerald J. M. Gazdar, and Roger M. Needham. 2000. Introduction: combining formal theories and statistical data in natural language processing. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 358(1769):1227–1238.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. In *arXiv preprint*. arXiv:1602.02410v2 [cs.CL], pages 1–11.

John D. Kelleher. 2016. Fundamentals of machine learning for neural machine translation. In *Proceedings of the Translating Europen Forum 2016: Focusing on Translation Technologies*. European Commission Directorate-General for Translation. https://doi.org/10.21427/D78012.

John D. Kelleher and Fintan J. Costello. 2005. Cognitive representations of projective prepositions. In *Proceedings of the Second ACL-SIGSEM workshop on the linguistic dimensions of prepositions and their use in computational linguistics formalisms and applications*. Association for Computational Linguistics, University of Essex, Colchester, United Kingdom, pages 119–127.

John D. Kelleher and Simon Dobnik. 2017. What is not where: the challenge of integrating spatial representations into deep learning architectures. In *CLASP Papers in Computational Linguistics: Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML 2017)*. Gothenburg, Sweden, volume 1, pages 41–52.

John D. Kelleher and Geert-Jan M. Kruijff. 2005a. A context-dependent algorithm for generating locative expressions in physically situated environments. In Graham Wilcock, Kristiina Jokinen, Chris Mellish, and Ehud Reiter, editors, *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*. Association for Computational Linguistics, Aberdeen, Scotland, pages 1–7.

John D. Kelleher and Geert-Jan M. Kruijff. 2005b. A context-dependent model of proximity in physically situated environments. In *Proceedings of the Second ACL-SIGSEM workshop on the linguistic dimensions of prepositions and their use in computational linguistics formalisms and applications*. Association for Computational Linguistics, University of Essex, Colchester, United Kingdom.

John D Kelleher, Brian Mac Namee, and Aoife D'Arcy. 2015. *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT Press.

John D. Kelleher, Robert Ross, Colm Sloan, and Brian Mac Namee. 2011. The effect of occlusion on the semantics of projective spatial terms: a case study in grounding language in perception. *Cognitive Processing* 12(1):95–108.

John D. Kelleher and Josef van Genabith. 2006. A computational model of the referential semantics of projective prepositions. In P. Saint-Dizier, editor, *Syntax and Semantics of Prepositions*, Kluwer Academic Publishers, Dordrecht, The Netherlands, Speech and Language Processing.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*. Phoenix, Arizona USA, pages 2741–2749.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123(1):32–73.

Geert-Jan M. Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I. Christensen. 2007. Situated dialogue and spatial organization: what, where... and why? *International Journal of Advanced Robotic Systems* 4(1):125–138.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2016. Multi-agent cooperation and the emergence of (natural) language. In *arXiv preprint*. arXiv:1612.07182v2 [cs.CL], pages 1–11.

Yann LeCun. 1989. Generalization and network design strategies. Technical report CRG-TR-89-4, Department of Computer Science, University of Toronto.

Gordon D. Logan and Daniel D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and Space*, MIT Press, Cambridge, MA, pages 493–530.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*. Lisbon, Portugal, pages 1412–1421.

Didier Maillat. 2003. *The semantics and pragmatics of directionals: a case study in English and French*. Ph.D. thesis, University of Oxford: Committee for Comparative Philology and General Linguistics, Oxford, United Kingdom.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. The MIT Press.

Dominic Massaro. 1988. Some criticisms of connectionist models of human performance. *Journal of Memory and Language* 27:213–234.

Ernan McMullin. 1968. What do physical models tell us? In Bob van Rootselaar and Johan Frederik Staal, editors, *Logic, Methodology and Science III: Proceedings of the Third International Congress for Logic, Methodology and Philosophy of Science, Amsterdam 1967*, North-Holland Publishing Company, pages 385–396.

Terry Regier. 1996. *The human semantic potential: Spatial language and constrained connectionism*. MIT Press.

Terry Regier and Laura A. Carlson. 2001. Grounding spatial language in perception: an empirical and computational investigation. *Journal of Experimental Psychology: General* 130(2):273–298.

Ardi Roelofs. 2014. A dorsal-pathway account of aphasic language production: The WEAVER++/ARC model. *Cortex* 59:33–48.

Giancarlo Salton, Robert Ross, and John D. Kelleher. 2017. Attentive language models. In *Proceedings of the 8th International Joing Conference on Natural Language Processing (IJCNLP)*. Taipei, Taiwan, pages 441–450.

Niels Schutte, Brian Mac Namee, and John D. Kelleher. 2017. Robot perception errors and human resolution strategies in situated human–robot dialogue. *Advanced Robotics* 31(5):243–257.

Stuart Shieber. 1986. *An Introduction to Unification-Based Approaches to Grammar*. CSLI Publications, Stanford.

Danijel Skočaj, Matej Kristan, Alen Vrečko, Marko Mahnič, Miroslav Janíček, Geert-Jan M. Kruijff, Marc Hanheide, Nick Hawes, Thomas Keller, Michael Zillich, and Kai Zhou. 2011. A system for interactive learning in dialogue with a tutor. In *IEEE/RSJ International Conference on Intelligent Robots and Systems IROS 2011*. San Francisco, CA, USA.

Luc Steels and Martin Loetzsch. 2009. Perspective alignment in spatial language. In Kenny R. Coventry, Thora Tenbrink, and John. A. Bateman, editors, *Spatial Language and Dialogue*, Oxford University Press.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, Curran Associates, Inc., pages 3104–3112.

Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research* 37(1):141–188.

Fjodor van Veen. 2016. The neural network ZOO. The Asimov Institute Blog posted on September 14. http://www.asimovinstitute.org/neural-network-zoo.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *arXiv preprint*. arXiv:1502.03044v3 [cs.LG], pages 1–22.