Semantic Web -1 (2022) 1–42
DOI 10.3233/SW-223228
IOS Press

Is neuro-symbolic AI meeting its promises in natural language processing? A structured review

Kyle Hamilton *, Aparna Nayak, Bojan Božić and Luca Longo

SFI Centre for Research Training in Machine Learning, School of Computer Science, Technological University Dublin, Republic of Ireland
E-mails: kyle.i.hamilton@mytudublin.ie, aparna.nayak@tudublin.ie, bojan.bozic@tudublin.ie, luca.longo@tudublin.ie

Editors: Monireh Ebrahimi, IBM, USA; Pascal Hitzler, Kansas State University, USA; Md Kamruzzaman Sarker, University of Hartford, USA; Daria Stepanova, Bosch Center for AI, Germany
Solicited reviews: Vered Shwartz, University of British Columbia, Canada; Filip Ilievski, USC Information Sciences Institute, USA; two anonymous reviewers

Abstract. Advocates for Neuro-Symbolic Artificial Intelligence (NeSy) assert that combining deep learning with symbolic reasoning will lead to stronger AI than either paradigm on its own. As successful as deep learning has been, it is generally accepted that even our best deep learning systems are not very good at abstract reasoning. And since reasoning is inextricably linked to language, it makes intuitive sense that Natural Language Processing (NLP), would be a particularly well-suited candidate for NeSy. We conduct a structured review of studies implementing NeSy for NLP, with the aim of answering the question of whether NeSy is indeed meeting its promises: reasoning, out-of-distribution generalization, interpretability, learning and reasoning from small data, and transferability to new domains. We examine the impact of knowledge representation, such as rules and semantic networks, language structure and relational structure, and whether implicit or explicit reasoning contributes to higher promise scores. We find that systems where logic is compiled into the neural network lead to the most NeSy goals being satisfied, while other factors such as knowledge representation, or type of neural architecture do not exhibit a clear correlation with goals being met. We find many discrepancies in how reasoning is defined, specifically in relation to human level reasoning, which impact decisions about model architectures and drive conclusions which are not always consistent across studies. Hence we advocate for a more methodical approach to the application of theories of human reasoning as well as the development of appropriate benchmarks, which we hope can lead to a better understanding of progress in the field. We make our data and code available on github for further analysis.[1]

Keywords: Neuro-symbolic artificial intelligence, natural language processing, deep learning, knowledge representation & reasoning, structured review

*Corresponding author. E-mail: kyle.i.hamilton@mytudublin.ie.
[1] https://github.com/kyleiwaniec/neuro-symbolic-ai-systematic-review

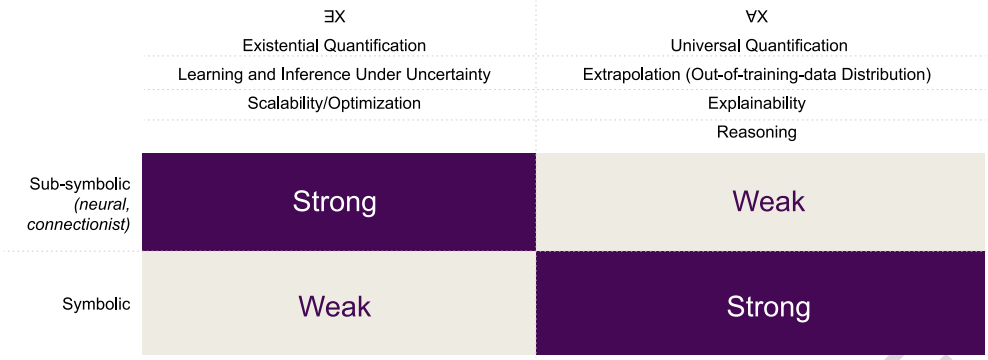


Fig. 1. Symbolic vs sub-symbolic strengths and weaknesses. Based on the work of Garcez et al. [50].

1. Introduction

At its core, Neuro-Symbolic AI (NeSy) is "the combination of deep learning and symbolic reasoning" [51]. The goal of NeSy is to address the weaknesses of each of symbolic and sub-symbolic (neural, connectionist) approaches while preserving their strengths (see Fig. 1). Thus NeSy promises to deliver a best-of-both-worlds approach which embodies the "two most fundamental aspects of intelligent cognitive behavior: the ability to learn from experience, and the ability to reason from what has been learned" [51,145].

Remarkable progress has been made on the learning side, especially in the area of Natural Language Processing (NLP) and in particular with deep learning architectures such as the Transformer [37,147]. However, these systems display certain intrinsic weaknesses which some researchers [104,113] argue cannot be addressed by deep learning alone and that in order to do even the most basic reasoning, we need rich representations which enable precise, human interpretable inference via mathematical logic.[2]

Recently, a discussion between Gary Marcus and Yoshua Bengio at the 2019 Montreal AI Debate prompted some passionate exchanges in AI circles, with Marcus arguing that "expecting a monolithic architecture to handle abstraction and reasoning is unrealistic", while Bengio defended the stance that "sequential reasoning can be performed while staying in a deep learning framework" [11]. Spurred by this discussion, and almost ironically, by the success of deep learning (and ergo, the clarity into its limitations), research into hybrid solutions has seen a dramatic increase – Fig. 2. At the same time, discussion in the AI community has culminated in "violent agreement" [80] that the next phase of AI research will be about "combining neural and symbolic approaches in the sense of NeSy AI [which] is at least a path forward to much stronger AI systems" [123]. Much of this discussion centers around the ability (or inability) of deep learning to reason, and in particular, to reason outside of the training distribution. Indeed, at IJCAI 2021, Yoshua Bengio affirms that "we need a new learning theory to deal with Out-of-Distribution generalization" [9]. Bengio's talk is titled "System 2 Deep Learning: Higher-Level Cognition, Agency, Out-of-Distribution Generalization and Causality." Here, System 2 refers to the System 1/System 2 dual process theory of human reasoning explicated by psychologist and Nobel laureate Daniel Kahneman in his 2011 book "Thinking, Fast and Slow" [77]. AI researchers [6,51,87,96,104,152,164] have drawn many parallels between the characteristics of sub-symbolic and symbolic AI systems and human reasoning with System 1/System 2. Broadly speaking, sub-symbolic (neural, deep-learning) architectures are said to be akin to the fast, intuitive, often biased and/or logically flawed System 1. And the more deliberative, slow, sequential System 2 can be thought of as symbolic or logical. But this is not the only theory of human reasoning as we will discuss later in this paper. It should also be noted that Kahneman himself has cautioned against the over reliance on the System 1/System 2 analogy in a followup discussion at the Montreal AI Debate 2 the following year, stating, "I think that this idea of two systems may have been adopted more than it should have been."[3]

[2] See also Besold et al. [12], p.17-18 for additional context.
[3] https://youtu.be/2zNd69ZGZ8o?t=161

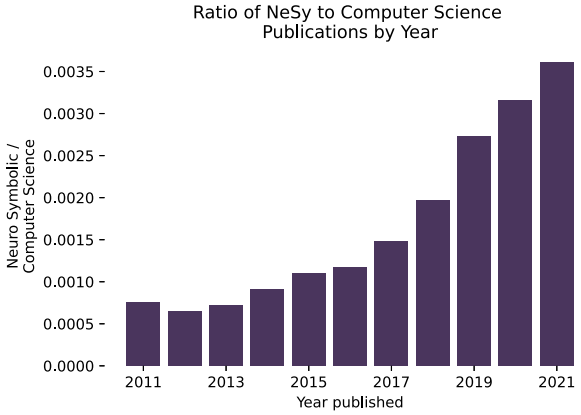


Fig. 2. Number of neuro symbolic articles published since 2010, normalized by the total number of all computer science articles published each year. The figure represents the unfiltered results from scopus given the search keywords described in Section 5.2.

1.1. Reasoning & language

"Language understanding in the broadest sense of the term, including question answering that requires commonsense reasoning, offers probably the most complete application area of neurosymbolic AI" [51]. This makes a lot of intuitive sense from a linguistic perspective. If we accept that language is compositional, with rules and structure, then it should be possible to obtain its meaning via logical reasoning. Compositionality in language was formalized by Richard Montague in the 1970s, in what is now referred to as *Montague grammar*: "The key idea is that compositionality requires the existence of a homomorphism between the expressions of a language and the meanings of those expressions."[4] In other words, there is a direct relationship between syntax and semantics (meaning). This is in line with Noam Chomsky's *Universal grammar*[5] which states that there is a structure to natural language which is innate and universal to all humans, and is governed by precise mathematical rules. While an analysis of the study of linguistics is beyond the scope of this paper, the key takeaway is this: what makes such theories so attractive to computational linguists is that meaning can be derived from syntactic structures which can be translated into computer programs. Today, industrial strength tools for extracting these structures (e.g., part-of-speech tagging, constituency parsing, dependency parsing) are readily available, such as for example NLTK[6] or SpaCy.[7] The challenge lies in representing and utilizing these structures in a way that both captures the semantics and is computationally efficient.

On the one hand, distributed representations are desirable because they can be efficiently processed by gradient descent (the backbone of deep learning). The downside is that the meaning embedded in a distributed representation is difficult if not impossible to decompose. So while a Large Language Model (LLM), a deep learning language model based on the principle of distributional semantics, may be very good at making certain types of predictions, it cannot be queried for answers not present in the training data by way of analogy or logic. We have also seen that even as these models get infeasibly large – the larger the model, the better the predictions [131] – they still fail on tasks requiring basic commonsense. The example in Fig. 3, given by Marcus and Davis in [105] is a case in point. In their now seminal paper *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, Bender et al. point to a wide variety of costs and risks associated with the rush for ever larger language models, including: environmental costs, financial costs, which in turn erect barriers to entry and limit who can contribute to this research area and which languages can benefit from the most advanced techniques; opportunity cost, as researchers pour effort away from directions requiring less resources; and the risk of substantial social harms due to the training data encoding

[4] https://plato.stanford.edu/entries/compositionality/#FormStat
[5] https://www.britannica.com/topic/universal-grammar
[6] https://www.nltk.org/
[7] https://spacy.io/

> You are having a small dinner party. You want to serve dinner in the living room. The dining room table is wider than the doorway, so to get it into the living room, you will have to **remove the door. You have a table saw, so you cut the door in half and remove the top half.**

Fig. 3. Third generation generative pre-trained transformer (GPT3) [20] text completion example. The prompt is rendered in regular font, while the GPT3 response is shown in bold. It is clear that GPT3 is incapable of commonsense.

hegemonic views that are harmful to marginalized populations, resulting in the amplification of existing biases, and the reinforcement of sexist, racist, ableist, etc. ideologies [8].

On the other hand, traditional symbolic approaches have also failed to capture the essence of human reasoning. While we may not yet understand exactly how people reason, it is generally accepted that human reasoning is nothing like the rigorous mathematical logic where the goal is validity. Though not for lack of ambition – Socrates got himself killed trying to get people to reason with logic [46]. In the *Dictionary of Cognitive Science* [43], Pascal Engel describes reasoning in a natural setting as "ridden with errors and paralogisms." Engel refers to Daniel Kahneman, Amos Tversky, Philip Wason, among others, who have conducted numerous experiments and written extensively showing how logical fallacies and "noise" can lead to those errors [77,78]. But even when the objective is not to emulate human thinking, but rather the execution of tasks which require precise, deterministic answers such as expert reasoning or planning, traditional symbolic reasoners are slow, cumbersome, and computationally intractable at scale, "typically subject to combinatorial explosions that limit both the number of axioms, the number of individuals and relations described by these axioms, and the depth of reasoning that is possible" [6]. For example, Description Logics (DLs) such as OWL[8] are used to reason over ontologies and knowledge graphs (KGs). However, one must accept a harsh trade-off between expressivity and complexity when choosing a DL flavor. Improving the performance of reasoning over ontologies and knowledge graphs that power search and information retrieval across the web is particularly relevant to the Semantic Web community. Hitzler et al. [65] report on recent research on neuro-symbolic integration in relation to the Semantic Web field, with a focus on the promises and possible benefits for both.

The remainder of this manuscript is structured as follows. Section 2 offers a brief history of NLP in the context of reasoning. Several recent surveys and their contributions to NeSy are discussed in Section 3, and are intended as an introduction to the field. Our contribution is given in Section 4, which also details the goals of NeSy selected for this survey. Section 5 describes the research methods employed for searching and analysing relevant studies. In Section 6 we analyze the results of the data extraction, how the studies reviewed fit into Henry Kautz's NeSy taxonomy [80], and we propose a simplified nomenclature for describing Kautz's NeSy categories. Section 7 discusses the limitations and challenges of the reviewed implementations. Section 8 presents limitations of this work and future directions for NeSy in NLP, followed by the conclusions in Section 9.

2. A brief history of NLP

The study of language and reasoning goes back thousands of years, but it was not until the 1960's that the first computational models were realized. The Association for Computational Linguistics (ACL)[9] was founded in 1962 for people working on computational problems involving human language, a field often referred to as either computational linguistics or Natural Language Processing (NLP). Common NLP tasks are illustrated in Fig. 4.

One of the first NLP projects was a chat-bot named ELIZA [155], written by Joseph Weizenbaum around 1965. Given a small hand crafted set of rules, ELIZA was able to hold an, albeit superficial, conversation, gaining tremendous popularity. Curiously, despite the program's simplicity those who interacted with it, attributed to it human-like emotions. These early systems were based on pattern matching and small rule-sets, and were very limited for obvious reasons. In the 1970s and 80s linguistically rich, logic-driven, grounded systems, largely influenced by Noam

[8] https://www.w3.org/2007/OWL/wiki/Direct_Semantics
[9] https://www.aclweb.org/portal/

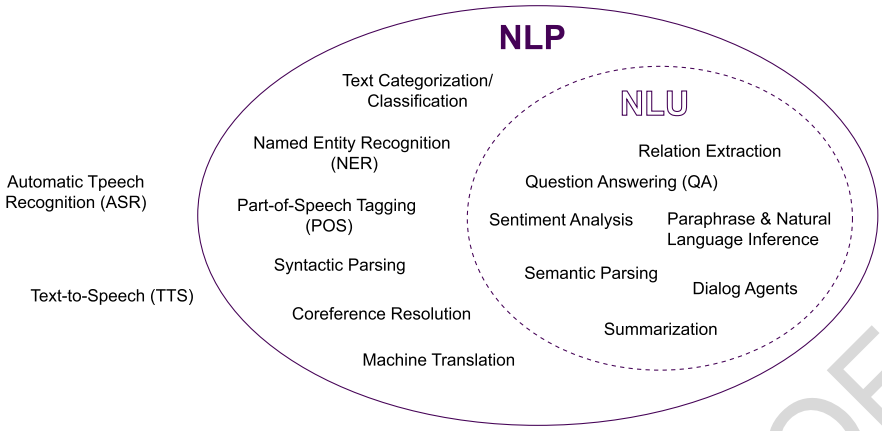


Fig. 4. Common natural language processing tasks [97].

Chomsky's *Universal Grammar*[10] were developed. The 1990s and early 2000s saw the 'statistical revolution' and the rise of machine learning, and work on NLP tasks focused on semantics, such as Natural Language Understanding (NLU), diminished for the next decade or so.[11] NLU returns to center stage, mixing techniques from previous years sometime around 2010. As a case in point, in 2011 IBM's Watson DeepQA computer system won first place on Jeopardy! for a prize of $1 million, competing against champions Brad Rutter and Ken Jennings.[12] DeepQA is a large ensemble of techniques and models, the vast majority of which was focused on general Information Retrieval (IR), NLP/NLU, Knowledge Representation & Reasoning (KRR), and Machine Learning (ML) [48]. Broadly speaking, DeepQA is a large neuro-symbolic question answering software pipeline. In the last decade, and especially in the last few years, the emphasis on deep learning has somewhat overshadowed traditional NLP approaches. The Long Short Term Memory (LSTM) [66] architecture paved the way for the Transformer, which has generated a huge amount of optimism leading some people to believe that "deep learning is going to be able to do everything."[13] However, as already mentioned, the success of the Transformer and Large Language Models (LLMs) has also served to highlight their inherent shortcomings. This brings us to the present, or the "3rd Wave" [51], which seeks to overcome those shortcomings by combining deep learning with symbolic reasoning and knowledge, and by integrating and expanding on the work of previous decades.

Areas of NLP which are said to benefit from this approach are ones which require some form of reasoning or logic. In particular, Natural Language Understanding (NLU), Natural Language Inference (NLI), and Natural Language Generation (NLG).

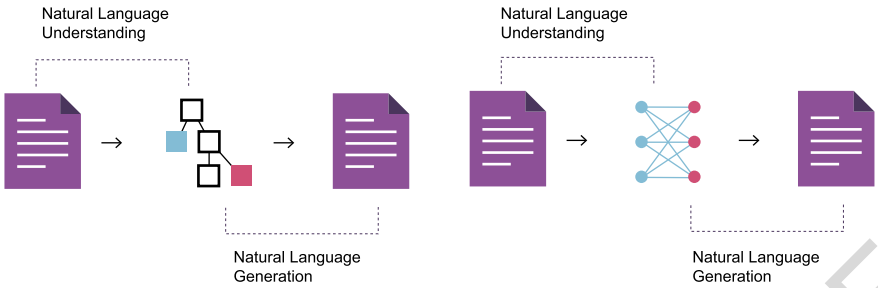
Natural Language Understanding (NLU) is a large subset of NLP containing topics particularly focused on semantics and meaning. The boundaries between NLP and NLU are not always clear and open to debate, and even when they are agreed upon, they're somewhat arbitrary, as it's a matter of convention and a reflection of history [97].

Natural Language Inference (NLI) enables tasks like semantic search, information retrieval, information extraction, machine translation, paraphrase acquisition, reading comprehension, and question answering. It is the problem of determining whether a natural language hypothesis h can reasonably be inferred from a given premise p [99]. For example, the premise "Hazel is an Australian Cattle Dog", entails the hypothesis "Hazel is a dog", and can be expressed in First Order Logic (FOL) by: $p \models h$.

Natural Language Generation (NLG) is the task of generating text or speech from non-linguistic (structured) input [55]. It can be seen as orthogonal to NLU, where the input is natural language. An end-to-end system can be made up of both NLU and NLG components. When that is the case, what happens in the middle is not always that

[10] https://www.cs.bham.ac.uk/~pjh/sem1a5/pt1/pt1_history.html
[11] https://nlp.stanford.edu/~wcmac/papers/20140716-UNLU.pdf
[12] https://www.youtube.com/watch?v=lI-M7O_bRNg
[13] https://www.technologyreview.com/2020/11/03/1011616/ai-godfather-geoffrey-hinton-deep-learning-will-do-everything/

clear-cut. A neural language model such as GPT3 [20] has no structured component, however, whether it performs "understanding" is subject to debate – Fig. 5.



(a) Symbolic view - reasoning is performed explicitly via rules and logic

(b) Connectionist view - reasoning is performed implicitly inside the neural network

Fig. 5. NLU takes as input unstructured text and produces output which can be reasoned over. NLG takes as input structured data and outputs a response in natural language.

3. Related work

Several recent surveys [6,12,50,51,87,123,152,163,164] cover neuro-symbolic architectures in detail. Our aim is not to produce another NeSy survey, but rather to examine whether the promises of NeSy in NLP are materializing. However, for completeness, and by way of introduction to the subject, we briefly summarize each of these surveys and provide references for the architectures under review.

In response to recent discussions in the AI community and the resurgence of interest in NeSy AI, Garcez et al. [51] synthesize the last 20 years of research in the field in the context of the aforementioned debate. The authors highlight the need for trustworthiness, interpretability, and accountability in AI systems, which ostensibly, NeSy is most suited to, in particular when it comes to natural language understanding. The authors also emphasize the distinction between commonsense knowledge and expert knowledge, and suggest that these two goals may ultimately lead to two distinct research directions: "those who seek to understand and model the brain, and those who seek to achieve or improve AI." Garcez at al. conclude that "Neurosymbolic AI is in need of standard benchmarks and associated comprehensibility tests which could in a principled way offer a fair comparative evaluation with other approaches" with a focus on the following goals: learning from fewer data, reasoning about extrapolation, reducing computational complexity, and reducing energy consumption[14] – Fig. 6.

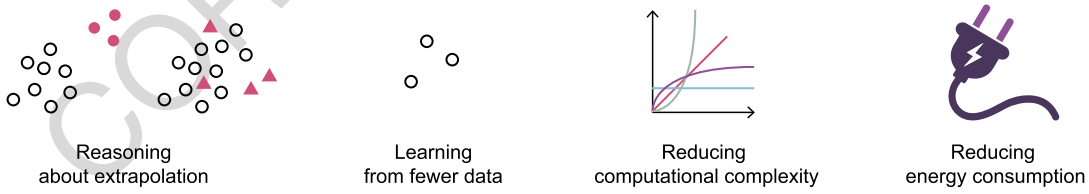


Fig. 6. Neuro-symbolic artificial intelligence promise areas [51].

Sarker et al. [123] survey recent work in the proceedings of leading AI conferences. The authors review a total of 43 papers and classify them according to Henry Kautz's categories,[15] as well as an earlier categorisation scheme

[14] Energy consumption is particularly significant when training Large Language Models which can cost in the thousands if not millions of dollars in electricity [131].

[15] Henry Kautz introduced a taxonomy of NeSy types at the Third AAAI Conference on AI [80]. We rely on this taxonomy to classify the studies under review, and discuss each type in detail in Section 6.2.3.

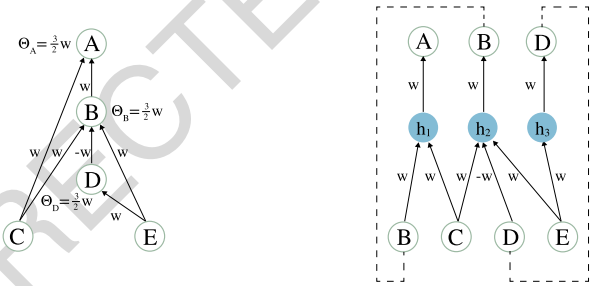
from 2005 [5]. Comparing the earlier research to the current trends, the authors confirm advancements on both the neural side, as well as the logic side, with a tendency towards more expressive logics being explored today than was thought tractable in the past, and the influence of the success of neural networks on the rise in interest in NeSy in general. Sarker et al. identify four areas of AI that can benefit from NeSy approaches: Learning from small data, out of distribution handling, intepretability, and error recovery – Fig. 7.

Fig. 7. Neuro-symbolic artificial intelligence promise areas [123].

The authors conclude that "more emphasis is needed, in the immediate future, on deepening the logical aspects in NeSy research even further, and to work towards a systematic understanding and toolbox for utilizing complex logics in this context." Based on the studies in our review, we come to a similar conclusion.

Garcez et al. [50] survey recent accomplishments for integrated machine learning and reasoning motivated by the need for interpretability and accountability in AI systems. According to [50], there are three main important features of a NeSy system: representation, extraction, and reasoning & learning. Symbolic knowledge can also be categorized into three groups: rule-based, formula-based, and embedding-based. The authors categorize and describe the following neuro-symbolic architectures.

Early systems such as KBANN [142] and CILP [53] embed propositional logic in a neural network by constraining the model parameters – Fig. 8.



(a) KBANN (Θ denotes a threshold). (b) CILP.

Fig. 8. Knowledge representation of $\phi = \{A \leftarrow B \wedge C, B \leftarrow C \wedge \neg D \wedge E, D \leftarrow E\}$ using KBANN and CILP. [50].

Tensorization is a process that embeds First Order Logic (FOL) symbols into real-valued tensors. Reasoning is performed through matrix computation. Examples include Logic Tensor Networks (LTNs) [129] and Neural Tensor Networks (NTNs) [136] – Fig. 9.

In **Neural-Symbolic Learning** the primary goal is learning, with the assistance of rules and logic. Different architectures are characterized by how the logic is incorporated into the network, and how it is translated into differentiable form.

– *Inductive Logic Programming* (ILP) [109] is a set of techniques for learning logic programs from examples:

  * Neural Logic Programming (NLP) [160]
  * Differentiable Inductive Logic Programming (∂ILP) [45]
  * Neural Theorem Prover (NTP) [120]
  * Neural Logic Machines (NLMs) [39]

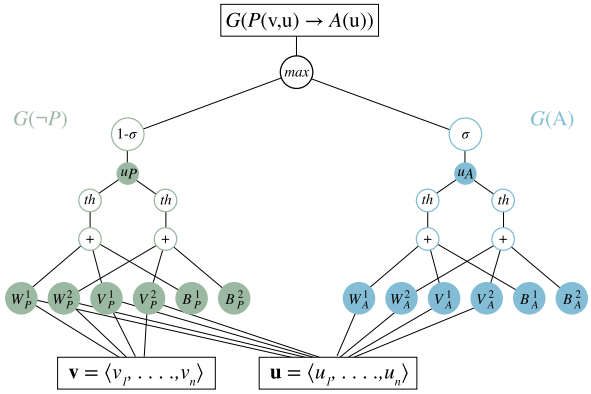


Fig. 9. Logic tensor network (LTN) for $P(x, y) \rightarrow A(y)$ with $G(x) = v$ and $G(y) = u$; G are grounding (vector representation) for symbols in first-order language [129].

- *Horizontal Hybrid Learning* combines expert knowledge in the form of rules/logic with data, thus are suitable to knowledge transfer learning (horizontally across domains).
- *Vertical Hybrid Learning* combines symbolic and sub-symbolic modules which take inspiration from neuroscience in that certain areas of the brain are responsible for processing input signals, while other areas perform logical thinking and reasoning (vertically for a single domain).

Neural-Symbolic Reasoning concerns itself with logical reasoning, as the name suggests, powered by neural computation. These consist of model-based, and theorem proving approaches. In early theorem proving systems such as SHRUTI [156] learning capability was limited. On the other hand, model-based approaches inside neural networks have been shown to demonstrate nonmonotonic, intuitionistic, abductive, and other forms of human reasoning capability. Hence, rather than attempting to perform both learning and reasoning in a single architecture, more recent designs tend to contain separate learning and reasoning modules which communicate with each other. The authors conclude that combining symbolic and sub-symbolic modules, in other words, the compositionality of neuro-symbolic systems, contributes to the development of explainable and accountable AI [150].

Yu et al. [163] divide neuro-symbolic systems into two types: heavy-reasoning light-learning and heavy-learning light-reasoning – Fig. 10. These are similar to the neural-symbolic reasoning and neural-symbolic learning categorization in [50] above. Heavy-reasoning light-learning mainly adopts the methods of the symbolic system to solve

Fig. 10. Two types of neuro-symbolic systems: heavy reasoning light learning, and heavy learning light reasoning [163].

the problem in machine reasoning, and introduces neural networks to assist in solving those problems, while heavy-learning light-reasoning mainly applies methods of the neural system to solve the problem in machine learning, and introduces symbolic knowledge in the training process.

- Heavy-reasoning light-learning (based on Statistical Relational Learning (SRL) [84])
  - ∗ Probabilistic Logic Programming (ProbLog) [34]

  - Markov Logic Network (MLN) [118]
  - Inductive Logic Programming (ILP) [109]

- Heavy-learning light-reasoning
  - *Regularization models* add symbols in the form of regular terms to the objective function as a kind of prior knowledge to guide training.
  - *Knowledge transfer models* integrate the knowledge graph that represents semantic information into the neural network model, making up for the lack of data by transferring semantic knowledge. Knowledge transfer models are mainly used to solve zero-shot learning and few-shot learning [154] tasks.

Besold et al. [12] examine neuro-symbolic learning and reasoning through the lens of cognitive science, cognitive neuro-science, and human-level artificial intelligence. This is a much more theoretical approach. The authors first describe some early systems such as CILP [53] and fibring, introduced by Garcez & Gabby [54]. Fibred networks work on the principle of recursion, where multiple neural networks are connected together, such that a fibring function in a network A, determines which neurons should be activated in a network B. A key characteristic of neuro-symbolic systems is modularity, where each network in the ensemble is responsible for a specific logic or task, increasing expressivity and allowing for non-classical logics to be represented such as connectionist modal, intuitionistic, temporal, nonmonotonic, epistemic and relational logic. Neuro-symbolic computation encompasses the integration of cognitive abilities – induction, deduction, abduction – and the study of mental models. The study of mental models has a long history, and the authors reference research from the field of neuro science and cognitive science, including the "binding" problem, dual process theory (e.g. System 1/System 2), and theories of affect; with the goal of formulating these in a neuro-symbolic system. Of particular interest to our work are the two sections on syntactic structures, and compositionality, as they both deal with modeling language. Psycho-linguists have different theories of language morphology (the study of the internal construction of words[16]), with some arguing for association based explanations (McClelland [76]), while others argue for a rule-based one (Pinker [115]) – the question being whether it is better to model language through a connectionist approach, per McClelland, or a symbolic one, as per Pinker. Whether to model language in a connectionist or symbolic manner hinges also on its inherent compositionality.[17]

Von Rueden et al. [152] propose a taxonomy for integrating prior knowledge into learning systems. This is an extensive work covering types of knowledge and knowledge representations, neuro-symbolic integration approaches, motivations for each approach, challenges and future directions. The authors categorize knowledge into three types: scientific knowledge, world knowledge, and expert knowledge. Furthermore, knowledge representations are classified into eight types – Fig. 11.

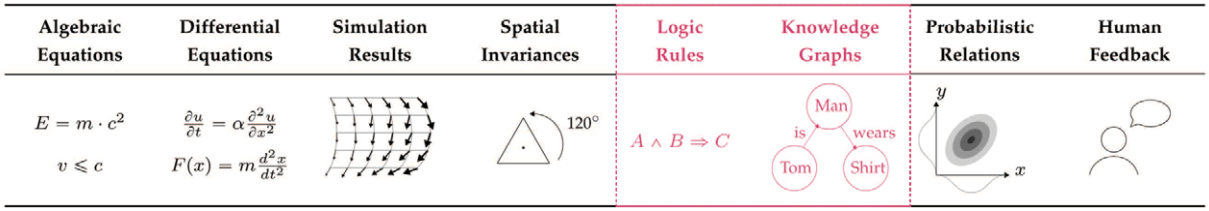


Fig. 11. Types of knowledge representation [152]. Given that our work deals with natural language as input, we are only concerned with logic rules (which we subdivide into rules and logic) and knowledge graphs (which we subdivide into frames and semantic networks) – see Section 6.2.2.

[16] https://www.britannica.com/topic/morphology-linguistics

[17] According to Noam Chomsky's theory of language, language is compositional, in the sense that a sentence is composed of phrases, which are in turn composed of sub-phrases, and so on, in a recursive manner. This idea enables the construction of infinite possibilities from finite means. This seems particularly well suited to a symbolic system which, given a finite set of rules should be capable of constructing/deconstructing, i.e., reasoning over, all possibilities. In contrast, a sub-symbolic, or distributional, system can never see the infinite amount of the data in the universe to learn from. For learning in infinite domains, see also [6]. https://www.britannica.com/biography/Noam-Chomsky/Rule-systems-in-Chomskyan-theories-of-language.

Zhang et al. [164] survey the area of neuro-symbolic reasoning on Knowledge Graphs (KGs). The authors contribute a unified reasoning framework for Knowledge Graph Completion (KGC) and Knowledge Graph Question Answering (KGQA). Among future directions, the authors advocate for taking inspiration from human cognition for neural-symbolic reasoning in KGs, alluding to the dual model of human reasoning (System 1/System 2). Additional future directions include:

- *Few-shot Reasoning* which addresses the issue of few labeled examples.
- *Reasoning upon Multi-sources* which incorporates additional information from unstructured text.
- *Dynamic Reasoning* which deals with inferring new facts evolving over time.
- *Analogical Reasoning (AR)* which involves the use of past experiences to solve problems that are similar to problems solved before. Case Based Reasoning (CBR) is an example of AR [138].
- *Knowledge Graph Pre-training* which enables transfer learning for domain adaptation.

Lamb et al. [87] review the state of the art on the use of Graph Neural Networks (GNNs) in NeSy – Fig. 12.

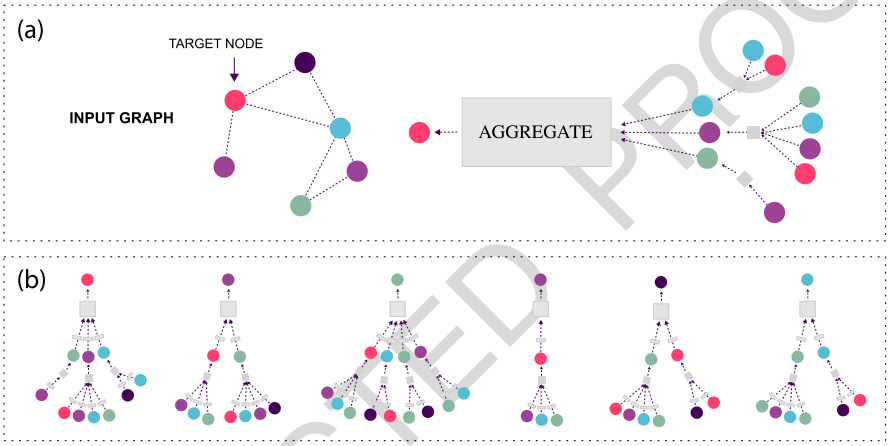


Fig. 12. Graph neural network (GNN) intuition: generate node embeddings based on local neighborhoods, where nodes aggregate information from their neighbors using neural networks (a). The network neighborhood defines a computation graph such that every node corresponds to a unique computation graph (b). The key distinctions are in how different approaches aggregate information across the layers [62].[18].

Similar to [51] and our work, this survey is motivated by the AI Debate in Montreal. Henry Kautz's NeSy taxonomy is used as a foundation for describing NeSy systems. A high level overview of state of the art neural architectures (convolutional layers, recurrent layers, and attention) is given, followed by a discussion of each of the following:

- Logic Tensor Networks (LTNs) [129] – Fig. 9.
- Pointer Networks [151]. Pointer networks are based on the encoder/decoder with attention (i.e. transformer) architecture, with the modification that the input length can vary. This architecture lends itself to combinatorial optimization problems such as the Traveling Salesperson Problem (TSP).
- Graph Convolutional Networks (GCNs) [81] can be thought of as a generalization of Convolutional Neural Networks (CNNs) for non-grid topologies.
- Graph Neural Network Model [125] – early GNN architecture similar to GCN.
- Message-passing Neural Networks – similar to GNN with a slightly modified update function [87].
- Graph Attention Networks (GATs) [148] – implement an attention mechanism enabling vertices to weigh neighbor representations during their aggregation. GATs are known to outperform typical GCN architectures for graph classification tasks.

[18] Tutorial slides associated with [62]: http://snap.stanford.edu/proj/embeddings-www/files/nrltutorial-part2-gnns.pdf.

According to the authors, GNNs endowed with attention mechanisms "are a promising direction of research towards the provision of rich reasoning and learning in [Kautz's] type 6 neuralsymbolic systems." In NLP, GATs have enabled substantial improvements in several tasks through transfer learning over pretrained transformer language models,[19] while GCNs have been shown to improve upon the state-of-the-art for seq2seq models [161]. GNN models have also been successfully applied to relational tasks over knowledge bases, such as link prediction [126].[20] The authors posit that the application of GNNs in NeSy will bring the following benefits:

- Extrapolation of a learned classification of graphs as Hamiltonian, to graphs of arbitrary size.
- Reasoning about a learned graph structure to generalise beyond the distribution of the training data.
- Reasoning about the $partOf(X; Y)$ relation (e.g., to make sense of handwritten MNIST digits and non-digits).
- Using an adequate self-attention mechanism to make combinatorial reasoning computationally efficient.

Belle [6] aims to disabuse the reader of the "common misconception that logic is for discrete properties, whereas probability theory and machine learning, more generally, is for continuous properties." The author advocates for tackling problems that symbolic logic and machine learning might struggle to address individually such as time, space, abstraction, causality, quantified generalizations, relational abstractions, unknown domains, and unforeseen examples.

Harmelen & Teije [146] present a conceptual framework to categorize the techniques for combining learning and reasoning via a set of design patterns. "Broadly recognized advantages of such design patterns are they distill previous experience in a reusable form for future design activities, they encourage re-use of code, they allow composition of such patterns into more complex systems, and they provide a common language in a community." A graphical notation is introduced where boxes with labels represent symbolic, and sub-symbolic modules, connected with arrows. Harmelen & Teije's boxology representation of AlphaGo is given in Fig. 13.

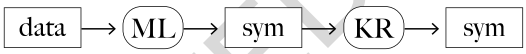


Fig. 13. Schematic diagram using the boxology graphical notation of the AlphaGo system. Ovals denote algorithmic components (i.e. objects that perform some computation), and boxes denote their input and output (i.e. data structures) [146].

Earlier surveys [5,13,49,52,63] tend to focus more on logic and logic programming, and less on learning, which is not surprising given that the ground breaking successes in deep learning are relatively recent. Several themes run through the above listed works, namely, the inherent strengths and weaknesses of symbolic and sub-symbolic techniques when taken in isolation, the types of problems which NeSy promises to solve, and the development of approaches over time.

Two future directions of particular interest to our work emerge: building systems which take inspiration from human cognition and reasoning, and the integration of unstructured data. To our knowledge there is no survey specifically covering the application of NeSy for Natural Language Processing (NLP) where the input data is both unstructured and replete with the ambiguities and inconsistencies of human reasoning.

4. Contributions

Our aim is to analyze recent work implementing NeSy in the language domain, to verify if the goals of NeSy are being realized, and to identify the challenges and future directions. We briefly describe each of the goals illustrated in Fig. 14, which we have identified based on our synthesis of the related work outlined above.

[19] The authors do not provide references to relevant works.

[20] While a detailed review of GNNs in NLP is beyond the scope of this work, we point the interested reader to an online resource dedicated to this topic: https://github.com/naganandy/graph-based-deep-learning-literature#computational-linguistics-conferences.

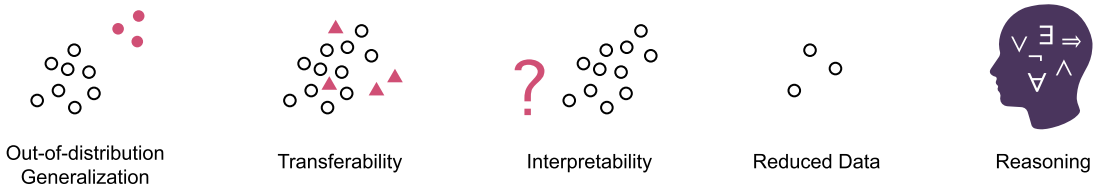


Fig. 14. Neuro-symbolic artificial intelligence goals.

4.1. Out-of-distribution (OOD) generalization

OOD generalization [132] refers to the ability of a model to extrapolate to phenomena not previously seen in the training data. The lack of OOD generalization in LLMs is often demonstrated by their inability perform commonsense reasoning, as in the example in Fig. 3.

4.2. Interpretability

As Machine Learning (ML) and AI become increasingly embedded in daily life, the need to hold ML/AI accountable is also growing. This is particularly true in sensitive domains such as healthcare, legal, and some business applications such as lending, where bias mitigation and fairness are critical. "An interpretable model is constrained, following a domain-specific set of constraints that make reasoning processes understandable" [121].

4.3. Reduced size of training data

State-of-the-art (SOTA) language models utilize massive amounts of data for training. This can cost in the thousands or even millions of dollars [131], take a very long time, and is neither environmentally friendly nor accessible to most researchers or businesses. The ability to learn from less data brings obvious benefits. But apart from the practical implications, there is something innately disappointing in LLMs' 'bigger hammer' approach. Science rewards parsimony and elegance, and NeSy promises to deliver results without the need for such massive scale. While this issue can be partially solved by fine tuning a pre-trained LLM using only a small amount labeled data, these techniques come with their own limitations. For example, Jiang et al. [74] discuss issues such as over-fitting the data of downstream tasks and forgetting the knowledge of the pre-trained model.

4.4. Transferability

Transferability is the ability of a model which was trained on one domain, to perform similarly well in a different domain. This can be particularly valuable, when the new domain has very few examples available for training. In such cases we might rely on knowledge transfer similar to the way a person might rely on abstract reasoning when faced with an unfamiliar situation [168].

4.5. Reasoning

According to Encyclopedia Britannica, "To reason is to draw inferences appropriate to the situation" [117]. Reasoning is not only a goal in its own right, but also the means by which the other above mentioned goals can be achieved. Not only is it one of the most difficult problems in AI,[21] it is one of the most contested. Also, a distinction must be made between human-level reasoning, or what is sometimes referred to as commonsense reasoning, and formal reasoning. While human-level reasoning can be ambiguous, error-prone, and difficult to specify, formal reasoning, or logic, follows strict rules and aims to be as precise as possible. The challenge lies in determining when it is appropriate to deploy one or the other or both, and how. In Section 7.1 we examine the uses of the term reasoning in more depth.

[21] As expressed by Luis Lamb at https://video.ibm.com/recorded/131288165.

5. Methods

Our review methodology is guided by the principles described in [82,111,112]. The data, queries, code, and additional details can be found in our github repository.[22]

5.1. Research questions

– Is Neuro-symbolic AI meeting its promises in NLP?

1. What are the existing studies on neurosymbolic AI (NeSy) in natural language processing (NLP)?
2. What are the current applications of NeSy in NLP?
3. How are symbolic and sub-symbolic techniques integrated and what are the advantages/disadvantages?

5.2. Search process

We chose Scopus to perform our initial search, as Scopus indexes most of the top journals and conferences we were interested in. In addition to Scopus, we searched the ACL Anthology database and the proceedings from conferences specific to Neuro-symbolic AI. It is possible we missed some relevant studies, but as our aim is to shed light on the field generally, our assumption is that these journals and proceedings are a good representation of the area as a whole. The included sources are listed in Appendix C. Since we were looking for studies which combine neural and symbolic approaches, our query consists of combinations of neural and symbolic terms as well as variations thereof, listed in Table 1. The keywords are deliberately broad, as it would be impossible to come up with a complete list of all possible keywords relevant to NeSy in NLP. More importantly, the focus of the work is not on specific subfields, each of which may warrant a review of its own, but rather on the explicit use of neuro-symbolic approaches regardless of subfield. Strictly speaking the only keywords that would cover this would be neuro-symbolic and its syntactic variants, but we relaxed this slightly on the basis that works which explore both symbolic reasoning and deep learning in combination (as per the definition in Section 1) may not necessarily have used the term neuro-symbolic.

Table 1
Search keywords

Neural terms	Symbolic terms	Neuro-symbolic terms
sub-symbolic	symbolic	neuro-symbolic
machine learning	reasoning	neural-symbolic
deep learning	logic	neuro symbolic
		neural symbolic
		neurosymbolic

The initial query was restricted to peer-reviewed English language journal articles and conference papers from the last 3 years, which produced a total of 21,462 results.

5.3. Study selection process

We further limit the Scopus articles to those published by the top 20 publishers as ranked by Scopus's CiteScore, which is based on number of citations normalized by the document count over a 4 year window,[23] and SJR (SCImago Journal Rank), a measure of prestige inspired by the PageRank algorithm over the citation network,[24] the union of which resulted in 29 publishers, and eliminated 19,560 studies, for a total of 1,519 journal articles and 383 conference papers for screening. Two researchers independently screened a sample of each of the 1,902 studies (articles and conference papers), based on the inclusion/exclusion criteria in Table 2. The selection process is illustrated in Fig. 15.

[22] https://github.com/kyleiwaniec/neuro-symbolic-ai-systematic-review
[23] https://service.elsevier.com/app/answers/detail/a_id/14880/kw/citescore/supporthub/scopus/
[24] https://service.elsevier.com/app/answers/detail/a_id/14883/supporthub/scopus/related/1/

Table 2
Inclusion/exclusion criteria

Inclusion	Exclusion
Input format: unstructured or semi structured text	Input format: structured query, images, speech, tabular data, categorical data, or any other data type which is not natural language text.
Output format: Any	Application: Theoretical Papers, Position Papers, Surveys, implementations of software pipelines from existing models
Application: Implementation of a novel architecture	The search keywords match, but the actual content does not
Language: English	Full text not available (Authors were contacted in these cases)

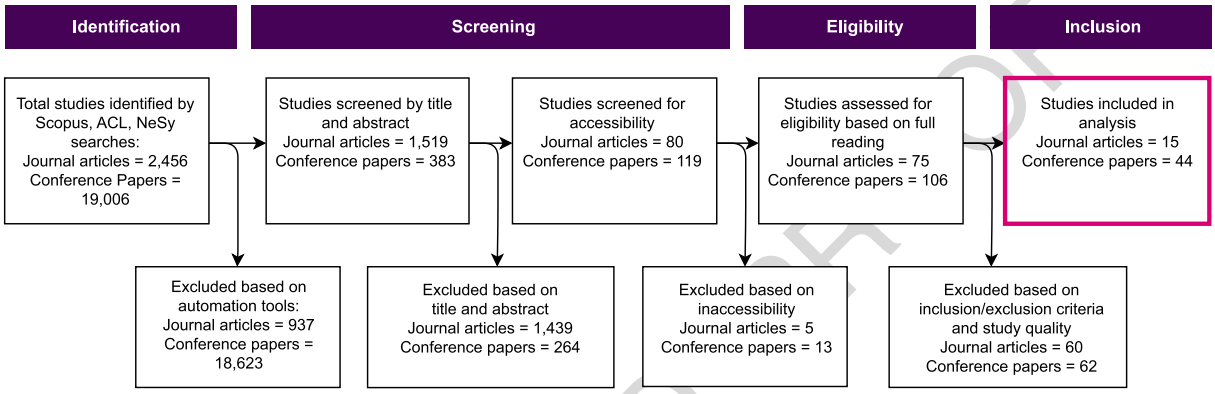


Fig. 15. Selection process diagram.

The inclusion criteria at this stage was intentionally broad, as the process itself was meant to be exploratory, and to inform the researchers of relevant topics within NeSy. As per best practices, this first round is also designed to understand and address inter-annotator disagreement. This unsurprisingly led to some researcher disagreement on inclusion, especially since studies need not have been explicitly labeled as neuro-symbolic to be classified as such. Agreement between researchers can be measured using the Cohen Kappa statistic, with values ranging from $[-1,1]$, where 0 represents the expected kappa score had the labels been assigned randomly, -1 indicates complete disagreement, and 1 indicates perfect agreement. Our score at this stage came to a modest 0.33. We observed that it was not always clear from the abstract alone whether the sub-symbolic and symbolic methods were integrated in a way that met the inclusion criteria.

To attain inter-annotator agreement and facilitate the next round of review, we kept a shared glossary of symbolic and sub-symbolic concepts as they presented themselves in the literature. We each reviewed all of the 1,902 studies, this time by way of a shallow reading of the full text of each study. Any disagreement at this stage was discussed in person with respect to the shared glossary. This process led to 75 journal articles and 106 conference papers marked for the final round of inclusion/exclusion.

5.4. Quality assessment

During the final round of inclusion/exclusion, the quality of each study was determined through the use of a nine-item questionnaire. Each of the following questions was answered with a binary value, and the study's quality was determined by calculating the ratio of positive answers. Less than a handful of studies were excluded due to a quality score of less than 50%.

1. Is there a clear and measurable research question?
2. Is the study put into context of other studies and research, and design decisions justified accordingly (number of references in the literature review/ introduction)?
3. Is it clearly stated in the study which other algorithms the study's algorithm(s) have been compared with?

4. Are the performance metrics used in the study explained and justified?
5. Is the analysis of the results relevant to the research question?
6. Does the test evidence support the findings presented?
7. Is the study algorithm sufficiently documented to be reproducible (independent researchers arriving at the same results using their own data and methods)?
8. Is code provided?
9. Are performance metrics provided (hardware, training time, inference time)?

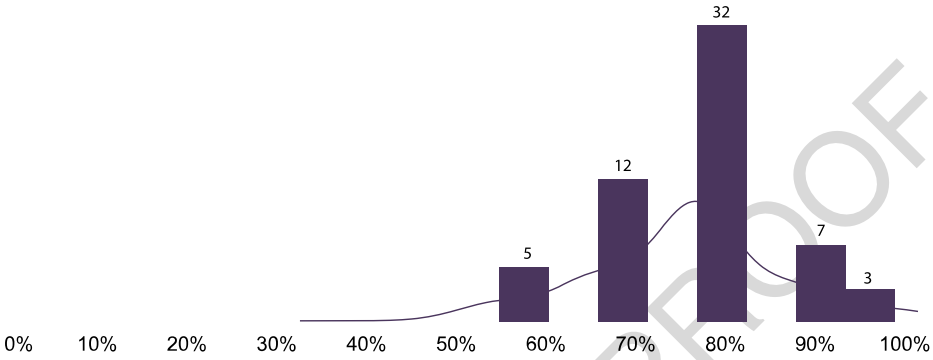
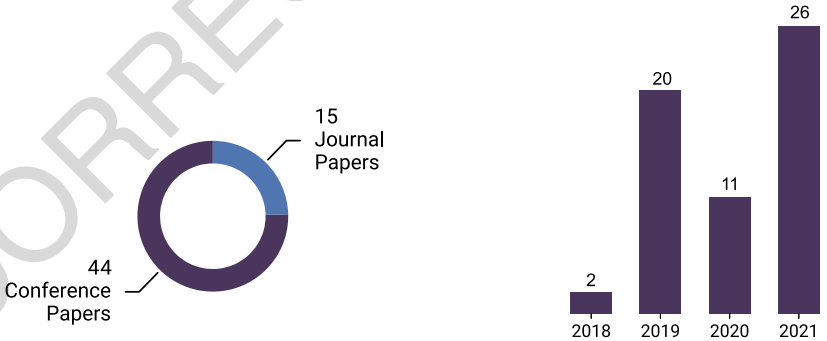


Fig. 16. Study quality.

More than 85% of the studies satisfy the requirements listed from Q1 to Q6. However, over 80% of the studies fail to provide source code or details related to the computing environment which makes the system difficult to reproduce. This leads to an overall reduction of the average quality score to 76.5% – Fig. 16.

Finally, a deep reading of each of the eligible studies led to 59 studies selected for inclusion. Data extraction was performed for each of the features outlined in Table 3. For acceptable values of individual features see Appendix B. The lists of neural and symbolic terms referenced in the table constitute the glossary items learned from conducting the selection process. Figure 17(a) shows the breakdown of conference papers vs journal articles, while Fig. 17(b) shows the number of studies published each year.



(a) Publication type (b) Published year

Fig. 17. Publications selected for inclusion.

6. Results, data analysis, taxonomies

We perform quantitative data analysis based on the extracted features in Table 3. Each study was labeled with terms from the aforementioned glossary, and each term in the glossary was classified as either symbolic, or neural.

Table 3
Data extraction features

Feature	Description
Business application	The stated objective or application of the proposed study. Often this is an NLP task, but this is not a requirement (e.g., "Medical decision support")
Technical application	Type of model output
Type of learning	Indicates learning method (supervised, unsupervised, etc.)
Knowledge representation	One of four categories: Rules, Logic, Frames, and Semantic networks
Type of reasoning	Indicates whether knowledge is represented implicitly (embedded) or explicitly (symbolic)
Language structure	Indicates whether linguistic structure is leveraged to facilitate reasoning
Relational structure	Indicates whether relational structure is leveraged to facilitate reasoning (e.g., part-of-speech tags, named entities, etc.)
Neural terms	List of neural architectures used by the models
Datasets	List of all datasets used for training and evaluation
Model description	Describes model architecture schematically
Evaluation Metrics	Evaluation metrics reported by the authors
Reported score	Model performance reported by the authors
Contribution	Novel contribution reported by the authors
Key-intake	Short description of the study
isNeSy	Indicates whether the authors label their study as Neuro-Symbolic
NeSy goals	For each of the goals listed in Section 1, indicates whether the goal is met as reported by the authors
Kautz category	List of categories from Kautz's taxonomy
NeSy category	List of categories from the proposed nomenclature
Study quality	Percentage of positive answers in the quality assessment questionnaire

A bi-product of this process are two taxonomies built bottom-up of concepts relevant to the set of studies under review. The two taxonomies are a reflection of the definition of NeSy provided earlier: "the combination of deep learning and symbolic reasoning." To make this definition more precise, we limit the type of combination that qualifies as neuro-symbolic. Specifically, the sub-symbolic and symbolic components must be integrated in a way such that one informs the other. By way of counter example, a system which is made up of two independent symbolic and sub-symbolic components would not be considered NeSy if there is no interaction between them. For example, while a system where one component is used to process one type of data, and the other is used to process another type of data may be an effective software pipeline design, we do not consider this type of solution neuro-symbolic as the two components do not interact in any way. Thus the definition becomes "the *integration* of deep learning and symbolic reasoning." It should be noted, that these terms are not always consistently defined in the literature. For example, in a much earlier survey, [5] split the interrelation (type of combination) of neuro-symbolic systems into *hybrid* and *integrated*, whereas we use the term *integrated* to cover both.

On the learning side, we have neural architectures (described in Section 6.2.1), and on the symbolic reasoning side we have knowledge representation (described in Section 6.2.2). These results are rendered in Table 4, with the addition of color representing a simple metric, or *promise score*, for each study. The promise score is simply the number of goals reported to have been satisfied by the solution in the study.

6.1. Exploratory data analysis

We plot the relationships between the features extracted from the studies, and the goals from Section 4 in an effort to identify any correlations between them, and ultimately to identify patterns leading to higher promise scores.

6.1.1. Business and technical applications

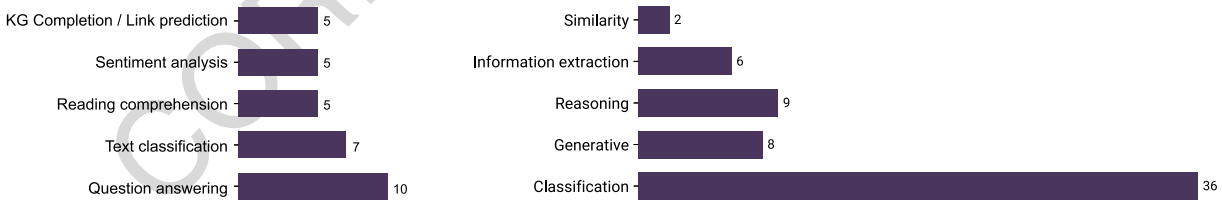
The *business application* is the stated application, or objective, of a given study. It is often but not always an NLP task, such as *text classification*, or *sentiment analysis*. It should be noted that in this example, sentiment analysis is a type of text classification, but while one author's stated objective is specific to sentiment, another author may be

interested in solving for text classification in general. As such there is no particular hierarchy or taxonomy associated with business applications. The relationship between all tasks, or business applications, and NeSy goals is shown in Fig. 18.

Fig. 18. Relationship between business applications and NeSy goals. Question answering is the most frequently occurring task, and is associated mainly with reasoning, reduced data, and to a lesser degree, interpretability.

The business application largely determines the type of model output, or what we term *technical application*. Most business applications are associated with a single (or at most two) technical applications. The exceptions being *question answering* and *reading comprehension*, which have been tackled as both inference and classification problems, or with the goal of information extraction or text generation. Question answering is the most frequently occurring task, and is associated mainly with reasoning, reduced data, and to a lesser degree, interpretability. On a philosophical level this seems somewhat disappointing, as one would hope that in receiving an answer, one could expect to understand why such an answer was given.

For completeness, the number of studies representing the technical applications and most frequently occurring business application is given in Fig. 19, while Fig. 20 illustrates the relationship between business applications, technical applications, and goals.



(a) Top Business Applications

(b) Technical Applications (model output)

Fig. 19. Number of studies in each application category.

6.1.2. *Type of learning*

Machine learning algorithms are classified as supervised, unsupervised, semi-supervised, curriculum or reinforcement learning, depending on the amount and type of supervision required during training [10,16,79]. Figure 21 demonstrates that the supervised method outnumbers all other approaches.

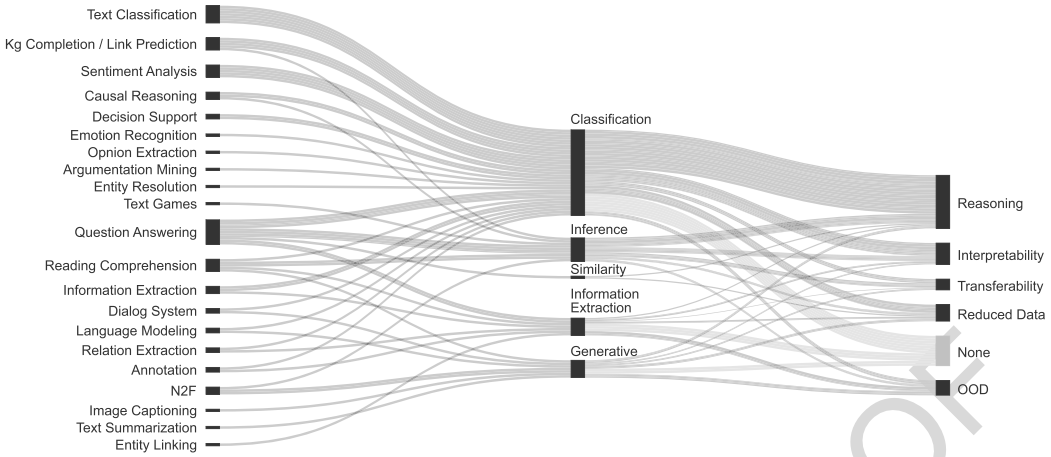


Fig. 20. Relationship between business applications, technical applications, and NeSy goals.

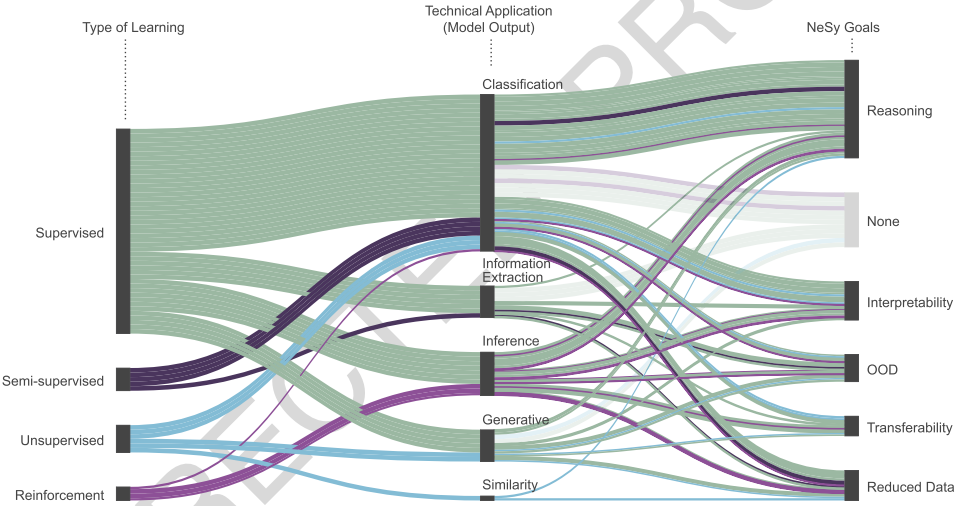


Fig. 21. Relationship between learning type, technical application, and NeSy goals. It is clear that supervised approaches dominate the field, are applied across a variety of technical applications, and there is no clear winner when it comes to goals.

6.1.3. Implicit vs explicit reasoning

The subset of tasks belonging to Natural Language Understanding (NLU) and Natural Language Generation (NLG) are often regarded as more difficult, and presumed to require reasoning. Given that reasoning was one of the keywords used for search, it is not surprising that many studies report reasoning as a characteristic of their model(s).

How reasoning is performed often depends on the underlying representation and what it facilitates. Sometimes the representations are obtained via explicit rules or logic, but are subsequently transformed into non-decomposable embeddings for learning. As such, we can say that any reasoning during the learning process is done implicitly. Studies utilizing Graph Neural Networks (GNNs) [26,59,70,91,124,165,167] would also be considered to be doing reasoning implicitly. The majority of the studies doing implicit reasoning leverage linguistic and/or relational structure to generate those internal representations. These studies meet 53 out of a possible 180 NeSy goals, where 180 = #goals * #studies, or 29.4%. For reasoning to be considered explicit, rules or logic must be applied during or after training. Studies which implement explicit reasoning perform slightly better, meeting 51 out of 135 goals, or 37.8% and generally require less training data. Additionally, 4 studies implement both implicit and explicit reasoning, at a NeSy promise rate of 40%. Of particular interest in this grouping is Bianchi et al. [14]'s implementation

of Logic Tensor Networks (LTNs), originally proposed by Serafini and Garcez in [130]. "LTNs can be used to do after-training reasoning over combinations of axioms which it was not trained on. Since LTNs are based on Neural Networks, they reach similar results while also achieving high explainability due to the fact that they ground first-order logic" [14]. Also in this grouping, Jiang et al. [75] propose a model where embeddings are learned by following the logic expressions encoded in huffman trees to represent deep first-order logic knowledge. Each node of the tree is a logic expression, thus hidden layers are interpretable.

Figure 22 shows the relationship between implicit & explicit reasoning and goals, while the relationship between knowledge representation, type of reasoning, and goals is shown in Fig. 23.

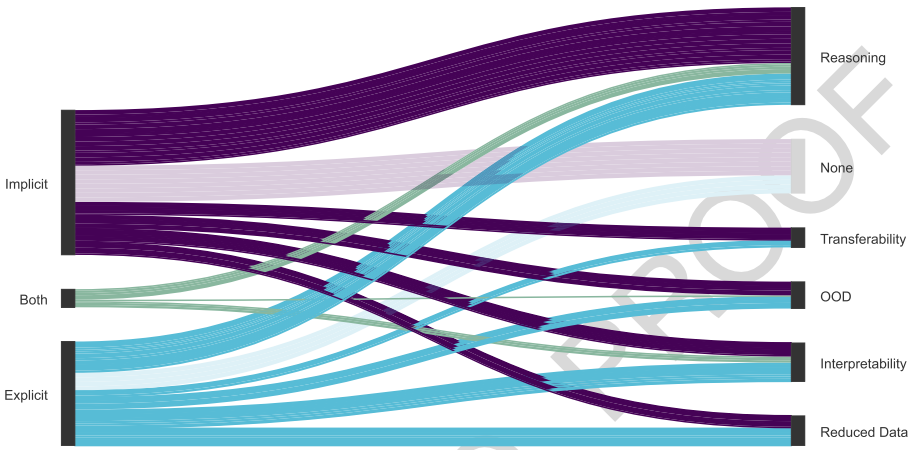


Fig. 22. Type of reasoning and goals. Around half, 48%, of studies where reasoning is performed explicitly mention interpretability as a feature. While nearly a third of studies performing reasoning implicitly do not meet any of the NeSy promises identified for this review.

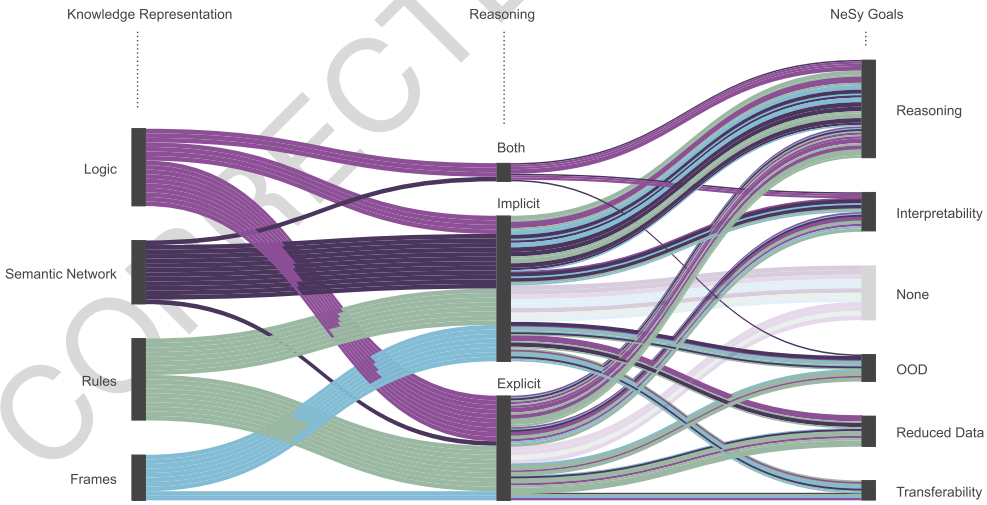


Fig. 23. Knowledge representation, type of reasoning, and goals. What is noteworthy, is that when semantic networks are utilized, reasoning is almost always done implicitly. The two exception are [14], and [165]. However, [14] utilizes FOL for explicit reasoning rather than its network component. On the other hand, [165] generate a novel interpretable reasoning graph as the output of their model.

6.1.4. Linguistic and relational structure

In the previous section we described how linguistic and relational structures can be leveraged to generate internal representations for the purpose of implicit reasoning. Here we plot the relationships between these structures and other extracted features and their interactions – Fig. 24.

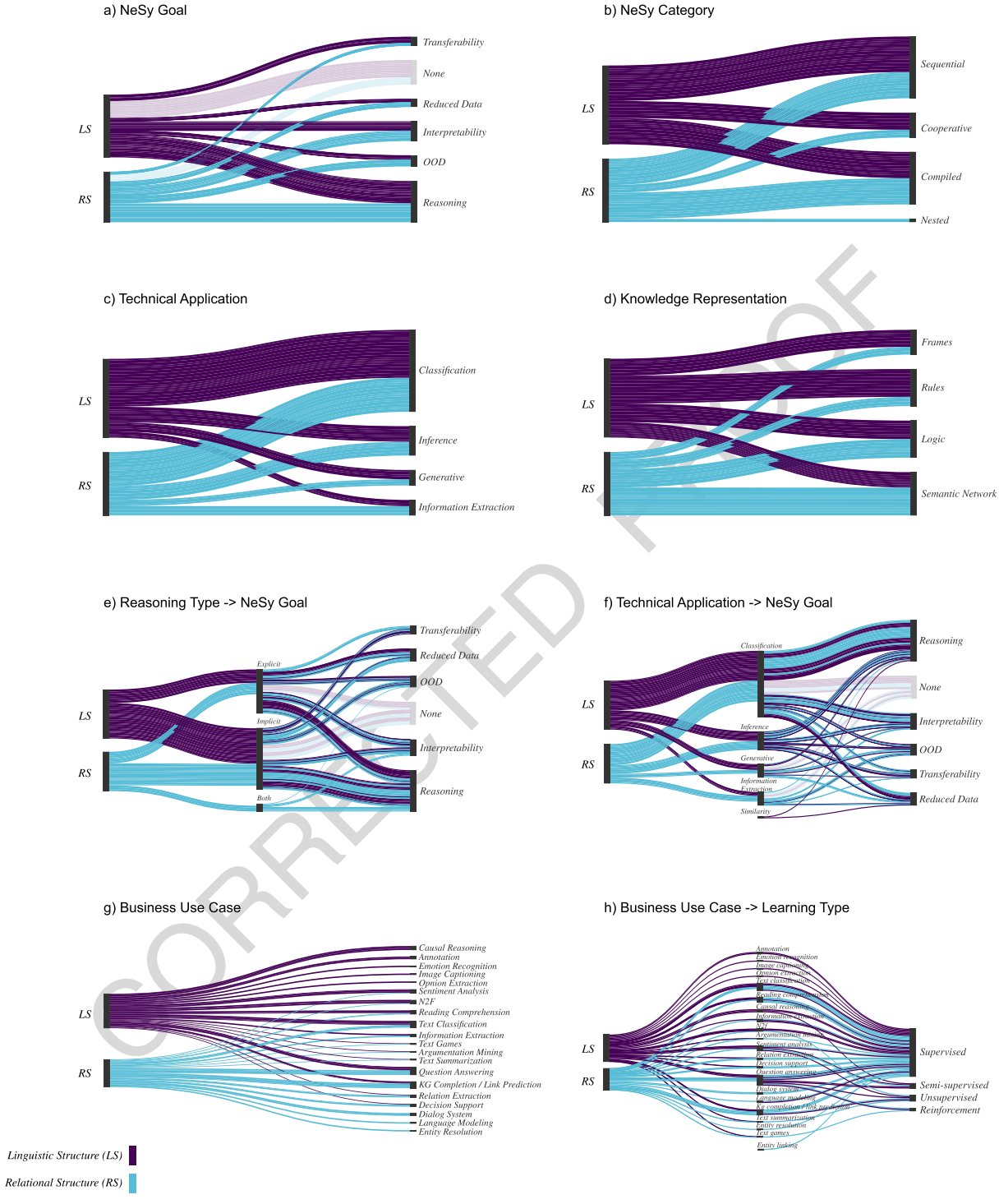


Fig. 24. Relationships between leveraged structures and extracted features. As can be seen in a), e), and f), studies leveraging linguistic structures often do not meet any NeSy goals, which runs counter to our original hypothesis. Further investigation into this phenomenon may be warranted. *note: studies which do no leverage either structure are not shown.*

Perhaps the most telling chart is the mapping between structures and goals, where many the studies leveraging linguistic structure do not meet any of the goals. This runs counter to the intuition that language is a natural fit for NeSy.

6.1.5. Datasets and benchmarks

Each study in our survey is based on a unique dataset, and a variety of metrics. Given that there are nearly as many business applications, or tasks, as there are studies, this is not surprising. As such it is not possible to compare the performance of the models reviewed. However, this brings up an interesting question, and that is how one might design a benchmark for NeSy in the first place. A discussion about benchmarks at the IBM Neuro-Symbolic AI Workshop 2022[25] resulted in general agreement that the most important characteristic of a good benchmark for NeSy is in the diversity of tasks tackled. Gary Marcus pointed out that current benchmarks can be solved extensionally, meaning they can be "gamed".[26] In other words, with enough attempts, a model can become very good at a specific task without solving the fundamental reasoning challenge. In essence, this is akin to over-fitting on the test set. The phenomenon can be exposed when adversarial examples are introduced such as described in [72], or through the observation that spurious correlations can be introduced in the annotation process as per [61]. This leads to models which are not able to generalize out of the training distribution. In contrast, to solve a task intensionally is to demonstrate "understanding" which is transferable to different tasks. This view is controversial with advocates of purely connectionist approaches arguing that "understanding" is not only ill defined, but also a moving target [51] – every time we solve for the current definition of understanding, the definition is revised to have to meet a higher bar. So instead of worrying about the semantics of "understanding", the panelists agreed that to make the benchmarks robust to gaming is to build in enormous variance in the types of tasks they tackle. Taking this a step further, Luis Lamb[27] proposed that instead of designing benchmarks for testing models, we should be designing challenges which encourage people to work on important real world problems. For a deeper dive, see the ACL-2021 Workshop on Benchmarking: Past, Present and Future (BPPF),[28] where some of the same issues pertaining specifically to NLP and NLU were discussed, as well as the challenges in interpreting performances across datasets, models, and with the evolution of language and context over time.

6.2. Taxonomies: Neural, symbolic, & neuro-symbolic

6.2.1. Neural

In the main, the extracted neural terms refer to the neural architecture implemented in a given study. We group these into higher level categories such as linear models, early generation (which includes CNNs), graphical models, sequence-to-sequence, and transformers. We include one study [134] which does not implement gradient descent, but rather Neuroevolution (NE). Neuroevolution involves genetic algorithms for learning neural network weights, topologies, or ensembles of networks by taking inspiration from biological nervous systems [90,108]. Neuroevolution is often employed in the service of Reinforcement Learning (RL). Studies which do not specify a particular architecture are categorised as Multilayer Perceptron (MLP) – Fig. 25.

We also include here neuro-symbolic architectures such as Logic Tensor Networks (LTN) [130], Recursive Neural Knowledge Networks (RNKN) [75], Tensor Product Representations (TPRs) [135], and Logical Neural Networks (LNN) [119] because they are suitable to optimization via gradient descent – Fig. 26.

6.2.2. Symbolic

The definition we adopted states that NeSy is *the integration of deep learning and symbolic reasoning*. Our neural taxonomy described above reflects the *deep learning* component. For the *symbolic reasoning* component we utilize four common Knowledge Representation (KR) categories: 1) production rules, 2) logical representation, 3) frames, and 4) semantic networks [7,18,33,92,137,143]. The following definitions are merely a glimpse at each of these topics, in order to provide a basic intuition.

[25] https://video.ibm.com/recorded/131288165.
[26] https://video.ibm.com/recorded/131288165 time-marker 43:00.
[27] https://video.ibm.com/recorded/131288165 time-marker 50:00.
[28] https://github.com/kwchurch/Benchmarking_past_present_future#S1

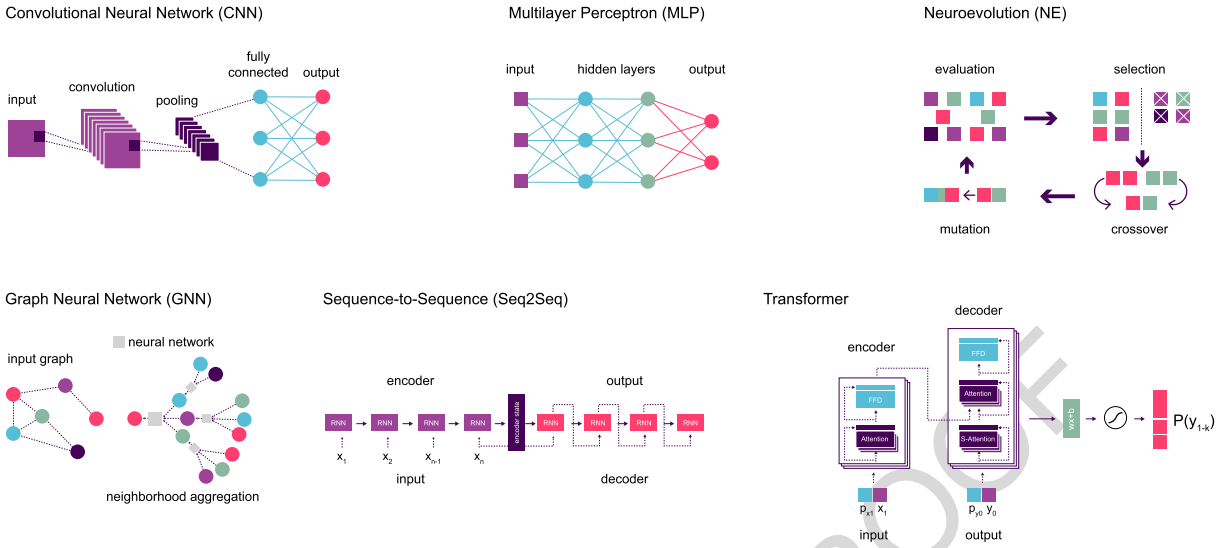


Fig. 25. Neural architectures represented in Table 4.

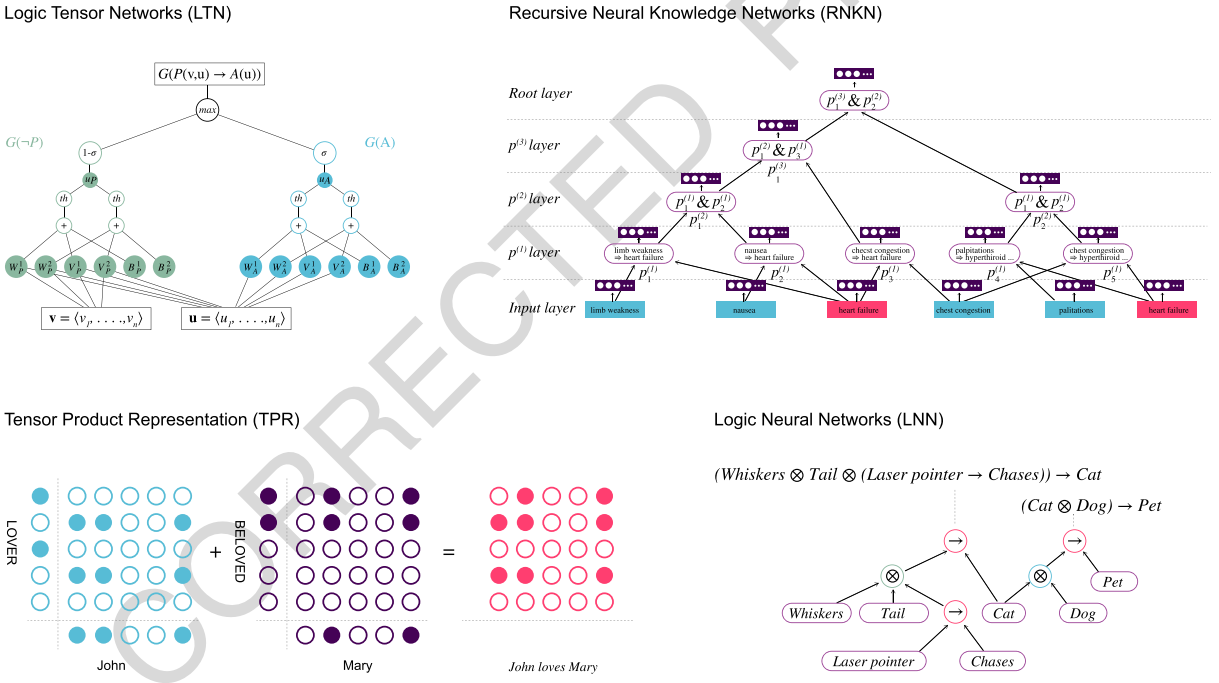


Fig. 26. Neuro-symbolic architectures represented in Table 4.

1. *Production rules* – A production rule is a two-part structure comprising an antecedent set of conditions and a consequent set of actions [18]. We usually write a rule in this form:

IF conditions THEN actions *ex) IF Bird THEN fly*

2. *Logical representation* – Logic is the study of entailment relations – languages, truth conditions, and rules of inference. [18,41]. A logic includes:
   - **Syntax**: specifies the symbols in the language and how they can be combined to form sentences. Hence facts about the world are represented as sentences in logic.
   - **Semantics**: specifies what facts in the world a sentence refers to. Hence, also specifies how you assign a truth value to a sentence based on its meaning in the world. A fact is a claim about the world, and may be true or false.
   - **Inference Procedure (reasoning)**: mechanical method for computing (deriving) new (true) sentences from existing sentences.

   The sentence "Not all birds can fly" in First Order Logic (FOL) looks like:

   $$\neg\big(\forall x Bird(x) \rightarrow Fly(x)\big)$$

   FOL is by no means the only choice, but as per [18] it is a simple and convenient one for the sake of illustration. Natural Logic (NL) for example, is a formal proof theory built on the syntax of human language, which can be traced to the syllogisms of Aristotle [21]. "For better or worse, most of the reasoning that is done in the world is done in natural language. And correspondingly, most uses of natural language involve reasoning of some sort. Thus it should not be too surprising to find that the logical structure that is necessary for natural language to be used as a tool for reasoning should correspond in some deep way to the grammatical structure of natural language" [86]. Implementations and extensions include [3,98,99,102]. Real-valued logics are often utilized in machine learning because they can be made differentiable and/or probabilistic [129] and were first introduced by Łukasiewicz at the turn of the 20th century [64,106]. Other, logic-based cognitive modelling approaches such as non-monotonic logic, attempt to deal with the complexities of human reasoning, epistemology, and defeasible inference [139].
3. *Frames* – Frames are objects which hold entities, their properties and methods. An individual frame schema looks like:

   $$\begin{aligned}&(Frame - name && (Penguin\\ &\quad \langle slot - name1\ filler1\rangle && \quad canFly:\ 0\\ &\quad \langle slot - name2\ filler2\rangle && \quad isA:\ \text{“}Bird''\\ &\quad ...) && \quad ...)\end{aligned}$$

   The frame and slot names are atomic symbols; the fillers are either atomic values (like numbers or strings) or the names of other individual frames [18]. This is similar to Object Oriented Programming (OOP), where the frame is analogous to the object, and slots and fillers are properties and values respectively.
4. *Semantic networks* – A semantic network is a structure for representing knowledge as a pattern of interconnected nodes and edges [137]. A frame network is a kind of semantic network where nodes are frames, and edges are the relationships between nodes. An example of a semantic network often used in NLU systems is WordNet[29] – a lexical database of English – Fig. 27. Today semantic networks are more often referred to as Knowledge Graphs (KGs).[30]

Table 4 shows which studies combine which of the above neural (6.2.1) and symbolic (6.2.2) categories as well as the number of NeSy goals satisfied.

[29] https://wordnet.princeton.edu/

[30] This term was popularized after Google introduced contextual information to search results from their semantic network under the brand name *Knowledge Graph* https://blog.google/products/search/introducing-knowledge-graph-things-not/.

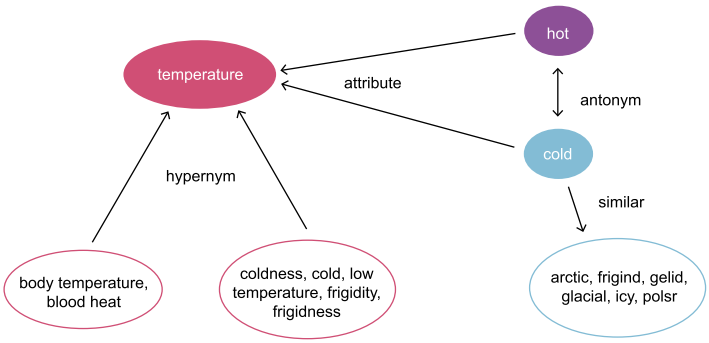


Fig. 27. English WordNet subgraph [107].

Table 4
Neural & symbolic combinations 1 2 3 4 5 number of NeSy goals satisfied out of the 5 described in Section 4. *Note: some studies use multiple techniques*

		Knowledge representation			
		Frames	Logic	Rules	Semantic net.
Linear Models	SVM			[40]	[71] [134]
Early Generation	MLP	[31] [30,158]	[17,44] [94]	[27] [162]	[167]
	CNN	[141]	[44] [23]	[4]	[56]
Graphical Models	DBN		[23]		
	GNN	[91]		[124] [70]	[26,167] [59] [165]
Sequence-to-Sequence	RNN	[19,141] [68] [25]	[58,60,89] [47] [23,127] [110]	[2] [1] [140] [36] [166] [116] [128] [103]	[56,85] [114] [95] [100]
	RcNN		[75]	[69]	
	w/Attn.		[93] [67]		[85]
Transformer		[29,83]	[157] [133,153] [67,110]	[159], [166] [32] [73]	[85] [35] [26,149] [167]
Neuro-Symbolic	LTN	[14]			
	RNKN		[75]		
	LNN		[24]	[73]	
	TPR	[25]		[69]	
Neuroevolution					[134]

6.2.3. *Neuro-symbolic*

NeSy systems can be categorized according to the nature of the combination of neural and symbolic techniques. At AAAI-20, Henry Kautz presented a taxonomy of 6 types of Neuro-Symbolic architectures with a brief example of each [80]. While Kautz has not provided any additional information beyond his talk at AAAI-20, several researchers have formed their own interpretations [51,87,123]. We have categorized all the reviewed studies according to Kautz's taxonomy as well as our proposed nomenclature – Fig. 28. Table 7 in Appendix A lists all the studies by category.

Type 1 *symbolic Neuro symbolic* is a special case where symbolic knowledge (such as words) is transformed into continuous vector space and thus encoded in the feature embeddings of an otherwise "standard" ML model. We opted to include these studies if the derived input features belong to the set of symbolic knowledge representations

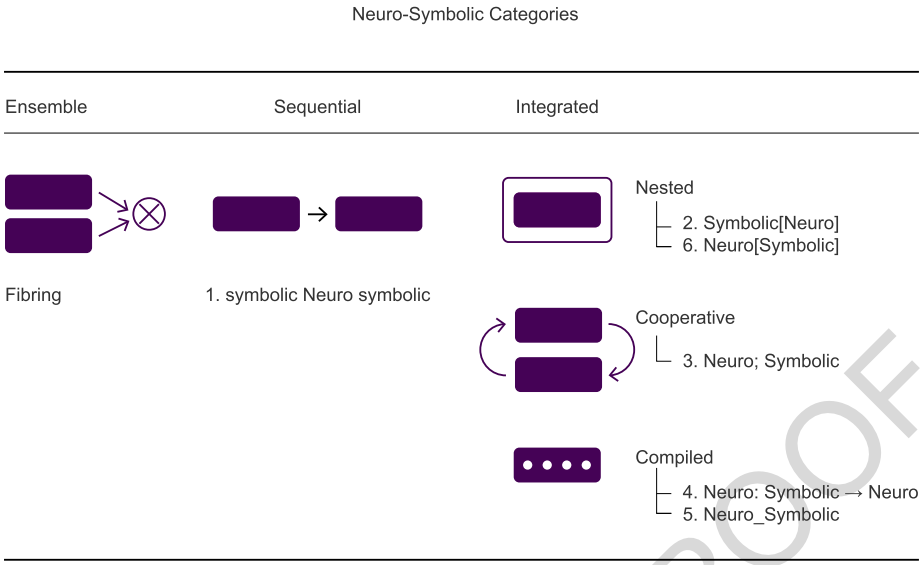


Fig. 28. Proposed neuro-symbolic artificial intelligence categories. Adapted from Henry Kautz.

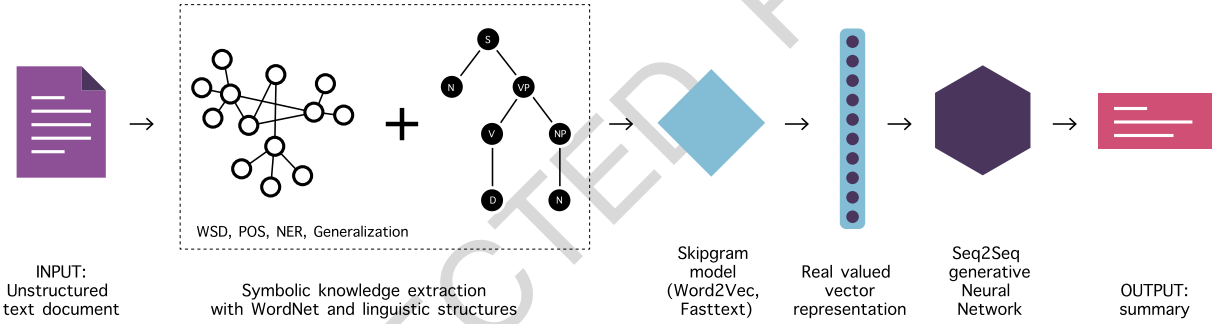


Fig. 29. Type 1 *sequential*. A symbolic knowledge representation module is used to generate rich embeddings for downstream machine learning [85].

described in Section 6.2 – Fig. 29. One could still argue that this is simply a case of good old fashioned feature engineering, and not particularly special, but we want to explore the idea that deep learning can perform reasoning, albeit implicitly, if provided with a rich knowledge representation in the pre-processing phase. We classify these studies as *Sequential*. Evaluating these studies as a group was particularly challenging as they have very little in common including different datasets, benchmarks and business applications. Half of the studies do not mention reasoning at all, and the ones that do are mainly executing rules on candidate solutions output by the neural models post hoc. In aggregate, only 26 out of a total of 115 (23 studies * 5 goals), or 22.6%, possible NeSy goals were met.

Type 2 *Symbolic[Neuro]* is what we describe as a *Nested* architecture, where a symbolic reasoning system is the primary system with neural components driving certain internal decisions. AlphaGo is the example given by Kautz, where the symbolic system is a Monte Carlo Tree Search with neural state estimators nominating next states. We found four studies that fit this architecture. We use [27] for the purposes of illustration – Fig. 30.

Type 3 *Neuro; Symbolic* is what we call *Cooperative*. Here, a neural network focuses on one task (e.g. object detection) and interacts via input/output with a symbolic reasoner specializing in a complementary task (e.g. query answering). Unstructured input is converted into symbolic representations which can be solved by a symbolic reasoner, which in turn informs the neural component which learns from the errors of the symbolic component. This process is iterated until convergence or a satisfactory output is produced. There are nine studies in this category, all but one of which utilize rules and/or logic for knowledge representation. A common theme among the coopera-

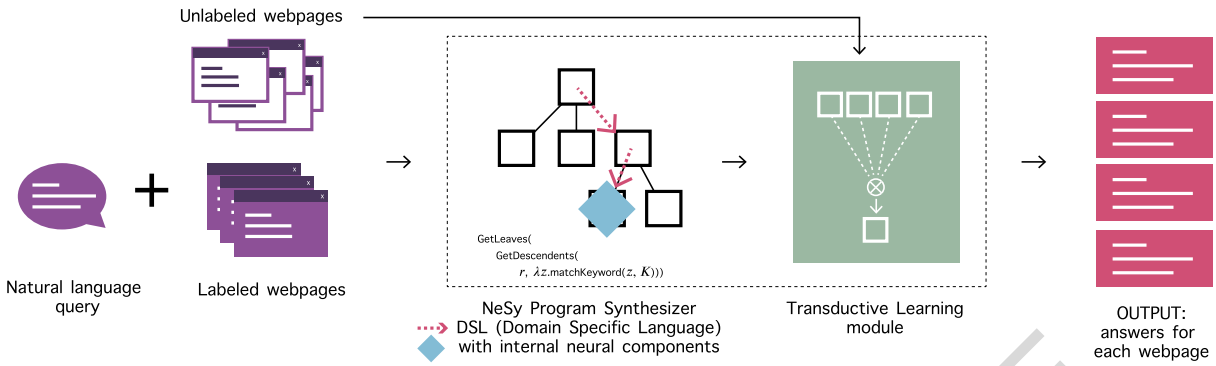


Fig. 30. Type 2 *nested*. Given a natural language query and a set of web pages, the system outputs answers for each page. A symbolic reasoner, which uses a custom domain specific language (DSL) to traverse the HTML, interacts with internal neural modules such as BERT which perform a number of natural language processing tasks. What is learned is a DSL program, using only a few labeled examples, which can generalize to a large number of heterogeneous web pages. The authors report large improvements in precision and recall scores over state-of-the art, in some cases over 50 points [27].

tive architectures is the business application of question answering. The Neuro-Symbolic Concept Learner (NS-CL) [103] – Fig. 31 – is an example of Type 3, meeting 4 out of the 5 NeSy goals. Its ability to perform well with reduced data is particularly impressive: "Using only 10% of the training images, our model is able to achieve comparable results with the baselines trained on the full dataset." Similarly, [162] report perfect performance on small datasets which they also attribute to the use of explicit and precise reasoning. Both studies display similar limitations, the use of synthetic datasets, and the need for handcrafted logic, a Domain Specific Language (DSL) in the case of [103], and Image Schemas in [162]. Six out of the nine studies leverage linguistic structures in some fashion, and in particular, [133] utilize *natural logic*, for a model which is both interpretable, and achieves state-of-the-art performance on two QA datasets. This work builds on [3,99].

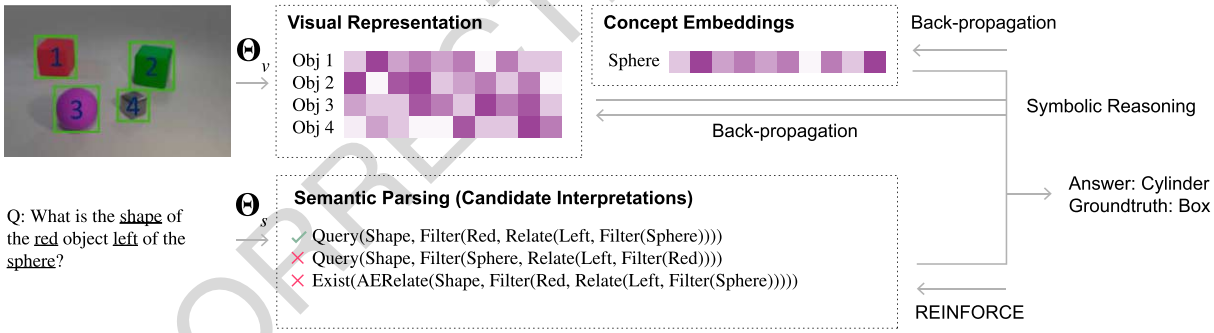


Fig. 31. Type 3 *cooperative*. The neuro-symbolic concept learner (NS-CL) jointly learns visual concepts, words, and semantic parsing of sentences without any explicit annotations. Given an input image, the visual perception module detects objects in the scene and extracts a deep, latent representation for each of them. The semantic parsing module translates an input question in natural language into an executable program given a domain specific language (DSL). The generated programs have a hierarchical structure of symbolic, functional modules, each fulfilling a specific operation over the scene representation. The explicit program semantics enjoys compositionality, interpretability, and generalizability [103].

Types 4 and 5, *Neuro: Symbolic → Neuro* and *Neuro_Symbolic* respectively, were originally presented by Kautz under one heading. After his presentation, Kautz modified the slide deck[31] separating these two types into systems where knowledge is compiled into the network weights, and where knowledge is compiled into the loss function. In Types 4 and 5, reasoning can be performed both implicitly and explicitly, in that it is calculated via gradient descent,

[31] https://henrykautz.com/talks/index.html.

but can also be performed post hoc. We have grouped studies belonging to these two categories under the moniker of *Compiled* systems, of which there are sixteen and seven respectively.

Deep Learning For Mathematics [88] is the canonical example of Type 4, where the input and output to the model are mathematical expressions. The model performs symbolic differentiation or integration, for example, given x^2 as input, the model outputs $2x$. The model exploits the tree structure of mathematical expressions, which are fed into a sequence-to-sequence architecture. This seems like a particularly fitting paradigm for natural language applications on the basis that structures such as parse trees can be similarly leveraged to output other meaningful structures such as for example: cause and effect relationships as exemplified in [166] and [159], or the generation of argument schemes as per [124]. The downside of many of these types of systems is the need for hand-crafted rules and logic [36,60,73,159]. In contrast, [24] learn rules from data (rule induction) by combining Logical Neural Networks (LNN)[32] with text-based Reinforcement Learning (RL). One could argue that this is a combination of Type 4, *compiled* (logic embedded in the network), and Type 3, *cooperative* (symbolic and sub-symbolic modules learning from each other in an iterative fashion). [24] is the only work we found which meets all five promises, and, it outperforms previous SOTA approaches – Fig. 32. Another example of a Type 4 system in our set of studies is

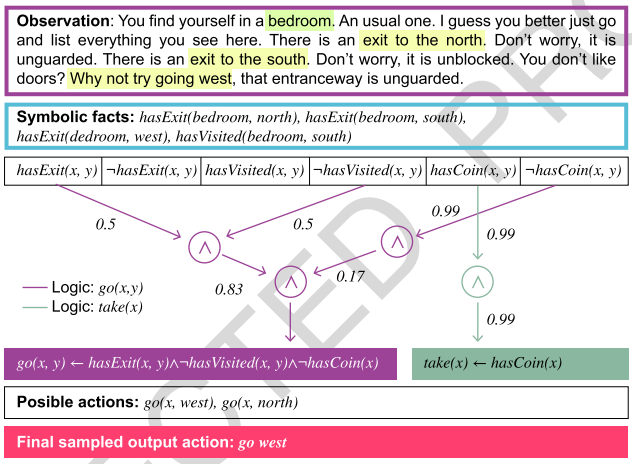


Fig. 32. Type 4 *compiled*. **S**ymbo**L**ic **A**ction policy for **T**extual **E**nvironments (**SLATE**) learns interpretable action policy for each action verb, *go* and *take*, from first-order symbolic states. The goal is to learn symbolic rules as logical connectives for generating action commands by gradient-based training [24].

proposed by [75]. Here, knowledge is encoded in the form of huffman trees made of triples and logic expressions, in order to jointly learn embeddings and model weights – Fig. 33. The model is intended for medical diagnosis decision support, where a requisite characteristic is interpretability, and this model meets that goal.

Type 5 comprises Tensor Product Representations (TPRs) [135], Logic Tensor Networks (LTNs) [129], Neural Tensor Networks (NTN) [136] and more broadly is referred to as tensorization, where logic acts as a constraint. LTN_{EE} [14] is an example of a *compiled* Type 5 system – Fig. 34.

Type 6 *Neuro[Symbolic]* is the most tightly integrated but perhaps the most elusive as there do not appear to be any recent implementations in existence. According to Kautz, this is the ultimate NeSy system which should be capable of efficient combinatorial reasoning at the level of super-intelligence, if not human intelligence.

Figure 35 shows the number of studies per category, and Fig. 36 illustrates the relationship between categories and goals. Table 5 shows the number of studies in each category per goal.

[32] It should be noted that the authors of the LNN classify their architecture as Type 7, which is explicitly outside of Kautz's 6 types. See Day 2, Session 2 of https://ibm.github.io/neuro-symbolic-ai/events/ns-summerschool2022/.

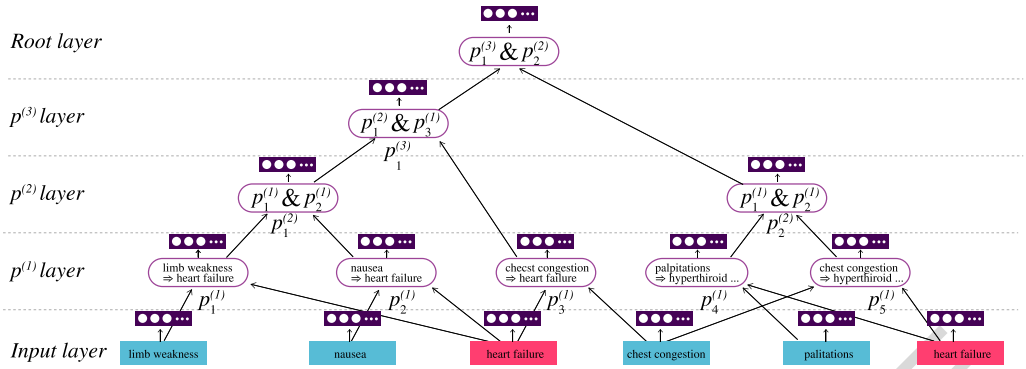


Fig. 33. Type 4 *compiled*. Huffman tree of the recursive neural knowledge network (RNKN), representing deep first-order logic knowledge. The first layer of the tree consists of entities, the second layer consists of relations ($x \rightarrow y$). Higher layers compute logic rules. The root node is the final embedding representing a document (in this case a single health record). Back propagation is used for optimization with softmax for calculating class probabilities [75].

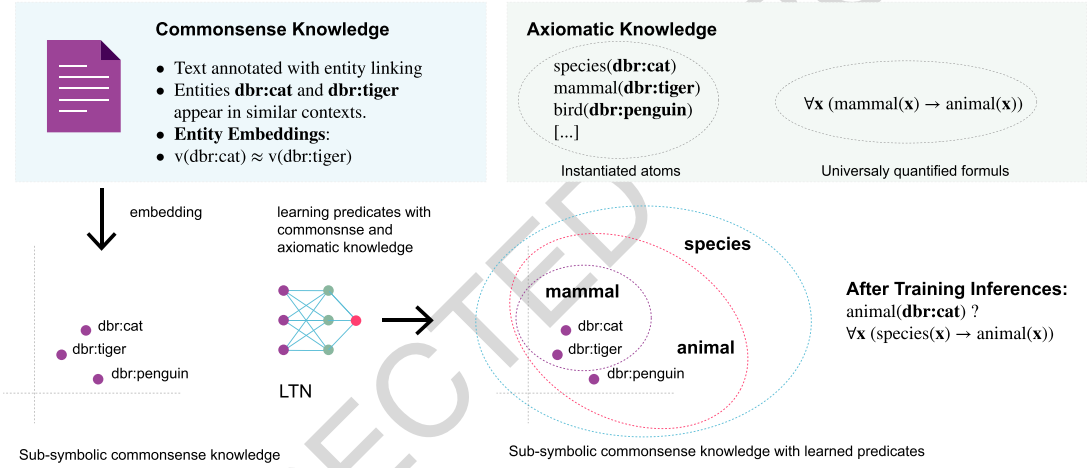
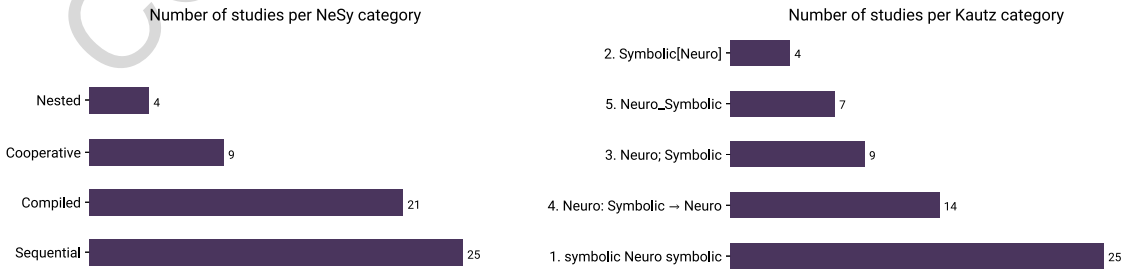


Fig. 34. Type 5 *compiled*. LTN_{EE} – using logic tensor networks (LTNs) it is possible to integrate axioms and facts (using first-order fuzzy logic to represent terms, functions, and predicates in a vector space) with commonsense knowledge represented in a sub-symbolic form (based on the principle of distributional semantics and implemented with Word2Vec) in one single model performing well in reasoning tasks. The major contribution of this work is to show that combining commonsense knowledge under the form of text-based entity embeddings with LTNs is not only simple, but it is also promising. LTNs can also be used to do after-training reasoning over combinations of axioms on which it was not trained [14].



(a) NeSy category (b) Kautz category

Fig. 35. Number of studies per category.

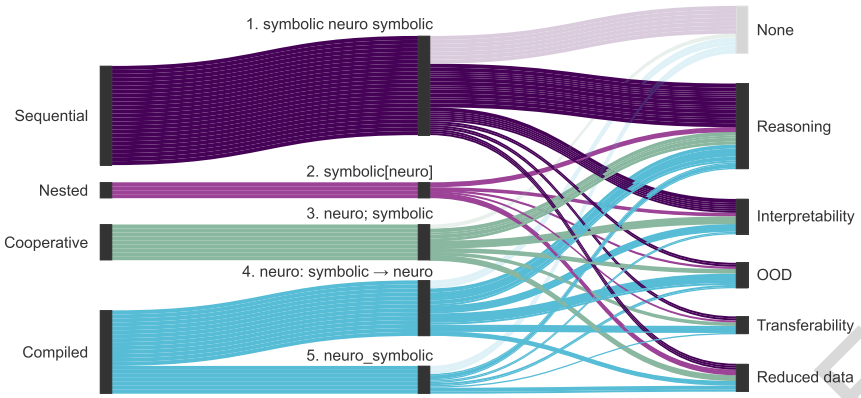


Fig. 36. NeSy categories to NeSy goals. There is no obvious pattern with respect to what types of goals are met within each of the NeSy categories.

Table 5

Number of studies meeting each goal. The *promise ratio* represents the percentage of goals reported to have been met out of the total number of possible goals (# of studies * 5 goals) in each category

	Compiled	Cooperative	Nested	Sequential
Reasoning	12	5	3	14
OOD	9	3	1	2
Interpretability	8	4	2	6
Reduced data	6	4	2	3
Transferability	7	2	1	2
Promise Ratio	29.5%	40%	45%	21.6%

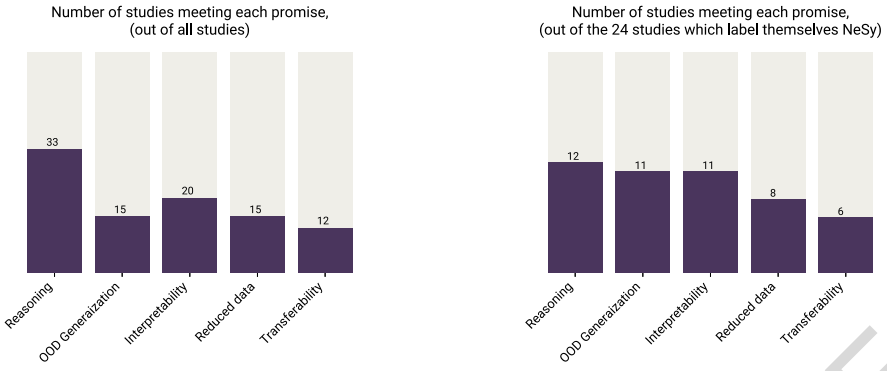
7. Discussion

All studies report performance either on par or above benchmarks, but we cannot compare studies based on performance as nearly every study uses a different dataset and benchmark as discussed in Section 6.1.5. Our focus is instead on whether the goals of NeSy are being met. Our *Promise Score* metric is not necessarily what the studies' authors were optimizing for or even reporting, especially studies which have not labeled themselves as NeSy per se. So we want to make it very clear that our analysis is not a judgement of the success of any particular study, but rather we seek to understand if the hypotheses about NeSy are materializing, namely that the combination of symbolic and sub-symbolic techniques will fulfill the goals described in Section 4: Out-of-distribution generalization, interpretability, tranferability, reduced data, and reasoning. The short answer is we are not there yet, as can be seen in Fig. 37. For a detailed breakdown of each goal and study see Table 6.

In Section 4.5 we put forward the hypothesis that reasoning is the means by which the other goals can be achieved. This is not evidenced in the studies we reviewed. Some possible explanations for this finding are: 1) The kind of reasoning required to fulfill the other goals is not the kind being implemented; 2) The approaches are theoretically promising, but the technical solutions need further development. Next we look at each of these possibilities.

7.1. Reasoning challenges

Thirty four out of the fifty nine studies mention reasoning as a characteristic of their solution. But there is a lot of variation in how reasoning is described and implemented. Given the overwhelming evidence of the fallibility of human reasoning, to understand language, AI researchers have sought guidance from disciplines such as psychology, cognitive linguistics, neuroscience, and philosophy. The challenge is that there are multiple competing theories of human reasoning and logic both across and within these disciplines. What we have discovered in our review, is a blurring of the lines between various types of logic, human reasoning, and mathematical reasoning, as well as



(a) All studies (b) NeSy studies only

Fig. 37. Proportion of studies which have met one or more of the 5 goals.

counter-productive assumptions about which theory to adopt. For example, drawing inspiration from "how people think", accepting that how people think is flawed, and subsequently attempting to build a model with a logical component, which by definition, is rooted in validity, seems counter productive to us. Although this does depend somewhat on the business application. For problems like MWP (*Math Word Problems*) [25,116,165], where answers are precise and unambiguous, less assumptions are needed. Additionally, the justification of "because that's how people think" is inconsistent. Some examples from the studies we reviewed include:

- [14] describe human reasoning in terms of a dual process of "subsymbolic commonsense" (strongly correlated with associative learning), and "axiomatic" knowledge (predicates and logic formulas) for structured inference.
- In [71] humans reason by way of analogy, and commonsense knowledge is represented in ConceptNet, a graphical representation of common concepts and their relationships.
- For [162] human reasoning can be modeled by Image Schemas (IS). Schemas are made up of logical rules on (Entity1, Relation, Entity2) tuples, such as transitivity, or inversion.
- [44] explain their choice of fuzzy logic for "its resemblance to human reasoning and natural language." This is a probabilistic approach which attempts to deal with uncertainty.
- [4] propose that human thought constructs can be modelled as cause-effect pairs. Commonsense is often described as the ability to draw causal conclusions from basic knowledge, for example: *If I drop the glass, it will break*.
- And [25] state that "when people perform explicit reasoning, they can typically describe the way to the conclusion step by step via relational descriptions."

But the most plausible hypothesis in our view is that of Schon et al. [127]: in order to emulate human reasoning, systems need to be flexible, be able to deal with contradicting evidence, evolving evidence, have access to enormous amounts of background knowledge, and include a combination of different techniques and logics. Most notably, no particular theory of reasoning is given. The argument put forward by Leslie Kaelbling at IBM Neuro-Symbolic AI Workshop 2022[33] is similarly appealing. Kaelbling points to the over-reliance on the System1/System2 analogy, and advocates for a much more diverse and dynamic approach. We posit that the type of reasoning employed should not be based solely on how we think people think, but on the attendant objective. This is in line with the "goal oriented" theory from neuroscience, in that reasoning involves many sub-systems: perception, information retrieval, decision making, planning, controlling, and executing, utilizing working memory, calculation, and pragmatics. But here the irony is not lost on us, and we acknowledge that by resorting to neuroscience for inspiration, we have just committed the same mischief for which we have been decrying our peers! But if we must resort to analogies with human reasoning then it is imperative to be as rigorous as possible. In their recent book, *A Formal Theory of Commonsense Psychology, How People Think People Think* [57], Gordon and Hobbs present a "large-scale logical

[33] https://researcher.watson.ibm.com/researcher/view_group.php?id=10897

Table 6
NeSy promises reported as having been met (y = yes, n = no)

Ref.	Score	Reasoning	OOD generalization	Interpretability	Reduced data	Transferability	isNeSy
[24]	5	y	y	y	y	y	y
[29,103]	4	y	y	y	y	n	y
[83]	4	y	y	y	n	y	y
[165]	4	y	n	y	y	y	n
[73,134]	4	n	y	y	y	y	y
[14,128]	3	y	y	y	n	n	y
[162]	3	y	n	y	y	n	n
[25]	3	y	n	y	n	y	n
[70]	3	n	y	n	y	y	n
[36]	2	y	y	n	n	n	y
[110,140]	2	y	n	y	n	n	y
[59,75,94,127]	2	y	n	y	n	n	n
[2,100]	2	y	n	n	y	n	n
[116,166]	2	y	n	n	n	y	y
[23]	2	y	n	n	n	y	n
[1]	2	n	y	n	y	n	y
[32,67]	2	n	y	n	y	n	n
[30]	2	n	y	n	n	y	n
[158]	2	n	n	y	n	y	n
[91,153]	1	y	n	n	n	n	y
[4,17,26,40,44,47,68,71,95,124,167]	1	y	n	n	n	n	n
[149]	1	n	y	n	n	n	y
[133]	1	n	n	y	n	n	y
[35]	1	n	n	y	n	n	n
[27,93]	1	n	n	n	y	n	y
[60,114,157,159]	0	n	n	n	n	n	y
[19,31,56,58,69,85,89,141]	0	n	n	n	n	n	n

formalization of commonsense psychology in support of humanlike artificial intelligence" to act as a baseline for researchers building intelligent AI systems. Santos et al. [122] take this a step in the direction we are advocating, by testing whether there is human annotator agreement when categorizing texts into Gordon and Hobbs' theories. "Our end-goal is to advocate for better design of commonsense benchmarks [and to] support the development of a formal logic for commonsense reasoning" [122]. It is difficult to imagine a single formal logic which would afford all of Gordon and Hobbs' 48 categories of reasoning tasks. Besold et al. [12] dedicate several pages to this topic under the heading of Neural-Symbolic Integration in and for Cognitive Science: Building Mental Models. In short, computational modelling of cognitive tasks and especially language processing is still considered a hard challenge.

7.2. *Technical challenges*

There is strong agreement that a successful NeSy system will be characterized by compositionality [6,12,15,22,28,50,51,144]. Compositionality allows for the construction of new meaning from learned building blocks thus enabling extrapolation beyond the training data distribution. To paraphrase Garcez et al., one should be able to query the trained network using a rich description language at an adequate level of abstraction [51]. The challenge is to come up with dense/compact differentialble representations while preserving the ability to decompose, or unbind, the learned representations for downstream reasoning tasks.

One such system, proposed by Bianchi et al. [14] is the LTN_{EE} – Fig. 34 – an extention of Logic Tensor Networks (LTNs), in which pre-trained embeddings are fed into the LTN. They show promising results on small datasets which have the important characteristic of being capable of after-training logical inferences. However, LTN_{EE} is limited by heavy computational requirements as the logic becomes more expressive, for example by the use of quantifiers.

Other studies [103,162] introduce logical inference within their solutions, but all require manually designed rules, and are limited by the domain expertise of the designer. Learning rules from data, or structure learning [42] is an ongoing research topic as pointed out by [152]. In [23] Chaturvedi et al. use fuzzy logic for emotion classification where explicit membership functions are learned. However, as stated by the authors, the classifier becomes very slow with the number of functions.

Compiled approaches involve translating logic into differentialble functions, which are either directly included as network nodes as in [75], or added as a constraint to the loss function, as in [38]. To achieve this, First Order Logic (FOL) can be operationalized using t-norms for example. To address the many types of reasoning as discussed in the previous section, we need to be able to incorporate other types of logic, such as temporal, modal, epistemic, non-monotonic, probabilistic, and more, which, presumably, are better able to model human reasoning.

In summary, formulating logic, or more broadly reasoning, in a differentiable fashion remains challenging.

8. Limitations & future work

We organized our analysis according to the characteristics extracted from the studies to test whether there were any patterns leading to NeSy goals. Another approach would be to reverse this perspective, and look at each goal separately to understand the characteristics leading to its fulfillment. However, each goal is really an entire field of study in and of itself, and we do not think we could have done justice to any of them by taking this approach. We spent a lot of time looking for signal in a very noisy environment where the studies we reviewed had very little in common. More can be said about what we did not find, than what we did. Another approach might be to narrow the criteria for the type of NLP task, while expanding the technical domain. In particular, a subset of tasks from the NLU domain could be a good starting point, as these tasks are often said to require reasoning.

We tried to be comprehensive in respect to the selected studies which led to the trade-off of less space dedicated to technical details or additional context from the neuro-symbolic discussion. There are a lot of ideas and concepts which we did not cover, such as, and in no particular order, Relational Statistical Learning (RSL), Inductive Logic Programming (ILP), DeepProbLog [101], Connectionist Modal Logics (CML), Extreme Learning Machines (ELM), Genetic Programming, grounding and proposinalization, Case Based Reasoning (CBR), Abstract Meaning Representation (AMR), to name but a few, some of which are covered in detail in other surveys [12,50].

Furthermore, we argued that we need differentiable forms of different types of logic, but we did not discuss how they might be implemented. A comprehensive point of reference such as this would be a very valuable contribution to the NeSy community, especially if the implementations were anchored in cognitive science and linguistics as discussed in 7.1.

Finally, the need for common datasets and benchmarks cannot be overstated.

9. Conclusion

We analyzed recent studies implementing NeSy for NLP in order to test whether the promises of NeSy are materializing in NLP. We attempted to find a pattern in a small and widely variable set of studies, and ultimately we do not believe there are enough results to draw definitive conclusions. Only 59 studies met the criteria for our review, and many of them (in the *Sequential* category) we would not consider truly integrated NeSy systems. The one thing studies which meet the most goals [24,29,73,83,103,134,165] have in common is that they all belong to the tightly integrated set of NeSy categories, *Cooperative* and *Compiled* which is good news for NeSy. Two out of these seven report lower computational cost than baselines, and performance on par or slightly above baselines, though we must reiterate that performance comparisons are not possible as discussed in Section 6.1.5. On the down side, we have seen that some studies suffer from high computational cost, and that explicit reasoning still often requires hand crafted domain specific rules and logic which makes them difficult to scale or generalize to other applications. Indeed, of the five goals, transferability to new domains was the least frequently satisfied.

Our view is that the lack of consensus around theories of reasoning and appropriate benchmarks is hindering our ability to evaluate progress. Hence we advocate for the development of robust reasoning theories and formal logics as well as the development of challenging benchmarks which not only measure the performance of specific implementations, but have the potential to address real world problems. Systems capable of capturing the nuances of natural language (i.e., ones that "understand" human reasoning) while returning sound conclusions (i.e., perform logical reasoning) could help combat some of the most consequential issues of our times such as mis- and dis-information, corporate propaganda such as climate change denialism, divisive political speech, and other harmful rhetoric in the social discourse.

Acknowledgements

This publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Appendix A. NeSy and Kautz categories

Table 7
NeSy and Kautz categories

NeSy (ours)	Kautz	Refs.
Sequential	1. symbolic Neuro symbolic	[2,4,17,19,26,30,31,35,40,44,47,56,59,67,68,85,89,94,100,114,127,140,141,158,167]
Nested	2. Symbolic[Neuro]	[23,27,29,110]
Cooperative	3. Neuro; Symbolic	[32,91,103,116,133,134,153,157,162]
Compiled	4. Neuro: Symbolic → Neuro	[24,36,60,70,73,75,83,95,124,128,149,159,165,166]
	5. Neuro_Symbolic	[1,14,25,58,69,71,93]

Appendix B. Allowed values

Table 8
Allowed values

Feature	Allowed values
Business application	Annotation, Argumentation mining, Causal Reasoning, Decision support, Dialog system, Emotion recognition, Entity Linking, Entity Resolution, Image captioning, Information extraction, KG Completion / link prediction, Language modeling, N2F, Opinion extraction, Question answering, Reading comprehension, Relation extraction, Sentiment analysis, Text classification, Text games, Text summarization
Technical application	Clustering, Generative, Inference, Classification, Information extraction, Similarity
Type of learning	Supervised, Unsupervised, Semi-supervised, Reinforcement, Curriculum
Type of reasoning	Implicit, Explicit, Both
Language structure	Yes, No
Relational structure	Yes, No
NeSy goals	Reasoning, OOD Generalization, Interpretability, Reduced data, Transferability
Kautz category	1. symbolic Neuro symbolic, 2. Symbolic[Neuro], 3. Neuro; Symbolic, 4. Neuro: Symbolic → Neuro, 5. Neuro_Symbolic, 6. Neuro[Symbolic]
NeSy category	Sequential, Nested, Cooperative, Compiled

Appendix C. Publishers

Table 9
Publishers included in the search

American Association for the Advancement of Science
American Chemical Society
American Institute of Physics
American Society for Microbiology
Association for Computing Machinery (ACM)
Association for Computational Linguistics (ACL)
Cairo University
Chongqing University of Posts and Telecommunications
Elsevier
Emerald
IEEE
IOS Press
Institute for Operations Research and the Management Sciences
King Saud University
MIT Press
Mary Ann Liebert
Morgan & Claypool Publishers
Now Publishers Inc
Optical Society of America
Oxford University Press
Public Library of Science
SAGE
Society for Industrial and Applied Mathematics
Springer Nature
Taylor & Francis
University of California Press
University of Minnesota
Wiley-Blackwell

Appendix D. Acronyms

Table 10
Acronyms and abbreviations

AAAI	Association for the Advancement of Artificial Intelligence
ACL	Association for Computational Linguistics
AI	Artificial Intelligence
AR	Analogical Reasoning
CBR	Case based reasoning
CNN	Convolutional Neural Network
DBN	Deep Belief Network
DL	Deep Learning

Table 10
(Continued)

DLs	Description Logic
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GNN	Graph Neural Network
GPT3	Third generation Generative Pre-trained Transformer
IJCAI	International Joint Conference on Artificial Intelligence
ILP	Inductive Logic Programming
KG	Knowledge Graphs
KGC	Knowledge Graph Completion
KGQA	Knowledge Graph Question Answering
KR	Knowledge Representation
KRR	Knowledge Representation & Reasoning
LNN	Logical Neural Networks
LLM	Large Language Models
LSTM	Long Short Term Memory
LTN	Logic Tenson Network
ML	Machine Learning
MLN	Markov Logic Network
MLP	Multilayer Perceptron
MWP	Math Word Problem
NE	Neuroevolution
NeSy	Neuro-Symbolic AI
NL	Natural Logic
NLI	Natural Language Inference
NLG	Natural Language Generation
NLM	Neural Logic Machine
NLP	Natural Language Processing
NLU	Natural Language Understanding
NN	Neural Network
NS-CL	Neuro-Symbolic Concept Learner
NTN	Neural Tensor Network
NTP	Neural Theorem Prover
OOD	Out-of-distribution
OOP	Object-oriented programming(paradigm)
OWL	Web Ontology Language
ProbLog	Probabilistic Logic Programming
RcNN	Recursive Neural Network
RL	Reinforcement Learning
RNKN	Recursive Neural Knowledge Network
RNN	Recurrent Neural Network
SOTA	State of the Art
SVM	Support Vector Machine
TPR	Tensor Product Representation
TSP	Traveling Salesperson Problem
(∂ILP)	Differentiable Inductive Logic Programming

References

[1] E. Altszyler, P. Brusco, N. Basiou, J. Byrnes and D. Vergyri, Zero-shot multi-domain dialog state tracking using prescriptive rules, in: *Proceedings of the 15th International Workshop on Neural-Symbolic Learning and Reasoning as Part of the 1st International Joint Conference on Learning & Reasoning (IJCLR 2021)*, Virtual Conference, October 25–27, 2021, A.S. d'Avila Garcez and E. Jiménez-Ruiz, eds, CEUR Workshop Proceedings, Vol. 2986, CEUR-WS.org, 2021, pp. 57–66.

[2] K. Amin, Cases without borders: Automating knowledge acquisition approach using deep autoencoders and Siamese networks in case-based reasoning, in: *31st IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2019*, Portland, OR, USA, November 4–6, 2019, IEEE, 2019, pp. 133–140. doi:10.1109/ICTAI.2019.00027.

[3] G. Angeli and C.D. Manning, NaturalLI: Natural logic inference for common sense reasoning, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang and W. Daelemans, eds, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 534–545. doi:10.3115/v1/d14-1059.

[4] R. Ayyanar, G. Koomullil and H. Ramasangu, Causal relation classification using convolutional neural networks and grammar tags, in: *2019 IEEE 16th India Council International Conference (INDICON)*, Rajkot, India, 2019, pp. 1–3. doi:10.1109/INDICON47234.2019.9028985.

[5] S. Bader and P. Hitzler, Dimensions of neural-symbolic integration — a structured survey, in: *We Will Show Them: Essays in Honour of Dov Gabbay*, S.Artemov, H. Barringer, A.S.D. Garcez, L.C. Lamb and J. Woods, eds, King's College Publications, 2005, pp. 167–194.

[6] V. Belle, Symbolic logic meets machine learning: A brief survey in infinite domains, in: *Scalable Uncertainty Management: 14th International Conference, SUM 2020, Proceedings*, Bozen-Bolzano, Italy, September 23–25, 2020, J. Davis and K. Tabia, eds, Springer-Verlag, Berlin, Heidelberg, 2020, pp. 3–16. ISBN 978-3-030-58449-8. doi:10.1007/978-3-030-58449-8_1.

[7] T.J.M. Bench-Capon, *Knowledge Representation: An Approach to Artificial Intelligence*, Academic Press Professional, Inc., USA, 1990. ISBN 0120864401.

[8] E.M. Bender, T. Gebru, A. McMillan-Major and S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT'21*, Association for Computing Machinery, New York, NY, USA, 2021, pp. 610–623. ISBN 9781450383097. doi:10.1145/3442188.3445922.

[9] Y. Bengio, System 2 Deep Learning: Higher-Level Cognition, Agency, Out-of-Distribution Generalization and Causality, 30th International Joint Conference on Artificial Intelligence, https://ijcai-21.org/invited-talks/, Accessed on 2022-06-04.

[10] Y. Bengio, J. Louradour, R. Collobert and J. Weston, Curriculum learning, in: *Proceedings of the 26th Annual International Conference on Machine Learning*, L. Bottou and M. Littman, eds, ICML'09, Association for Computing Machinery, New York, NY, USA, 2009, pp. 41–48. ISBN 9781605585161. doi:10.1145/1553374.1553380.

[11] Y. Bengio, G. Marcus, V. Boucher, AI DEBATE! Yoshua Bengio vs Gary Marcus, Montreal.AI, https://montrealartificialintelligence.com/aidebate/, Accessed on 2022-06-04.

[12] T.R. Besold, A.D. Garcez, S. Bader, H. Bowman, P. Domingos, P. Hitzler, K.-U. Kuehnberger, L.C. Lamb, D. Lowd, P.M.V. Lima, L. de Penning, G. Pinkas, H. Poon and G. Zaverucha, Neural-Symbolic Learning and Reasoning: A Survey and Interpretation, 2017, arXiv:1711.03902. doi:10.48550/ARXIV.1711.03902.

[13] T.R. Besold and K.-U. Kühnberger, Towards integrated neural-symbolic systems for human-level AI: Two research programs helping to bridge the gaps, *Biologically Inspired Cognitive Architectures* **14** (2015), 97–110. doi:10.1016/j.bica.2015.09.003.

[14] F. Bianchi, M. Palmonari, P. Hitzler and L. Serafini, in: *Complementing Logical Reasoning with Sub-Symbolic Commonsense*, Lecture Notes in Computer Science 11784 LNCS, 2019, pp. 161–170. doi:10.1007/978-3-030-31095-0_11.

[15] G. Boleda, Distributional semantics and linguistic theory, *Annual Review of Linguistics* **6**(1) (2020), 213–234. doi:10.1146/annurev-linguistics-011619-030303.

[16] G. Bonaccorso, *Machine Learning Algorithms*, Packt Publishing Ltd, 2017.

[17] M. Bounabi, K. Elmoutaouakil and K. Satori, A new neutrosophic TF-IDF term weighting for text mining tasks: Text classification use case, *International Journal of Web Information Systems* **17**(3) (2021), 229–249. doi:10.1108/IJWIS-11-2020-0067.

[18] R.J. Brachman and H.J. Levesque, *Knowledge Representation and Reasoning*, Elsevier, 2004. ISBN 978-1-55860-932-7.

[19] A.M.P. Braşoveanu and R. Andonie, Semantic fake news detection: A machine learning perspective, in: *Advances in Computational Intelligence*, I. Rojas, G. Joya and A. Catala, eds, Springer International Publishing, Cham, 2019, pp. 656–667. ISBN 978-3-030-20521-8. doi:10.1007/978-3-030-20521-8_54.

[20] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, Language Models are Few-Shot Learners, 2020, arXiv:2005.14165. doi:10.48550/ARXIV.2005.14165.

[21] J. Byszuk, M. Woźniak, M. Kestemont, A. Leśniak, W. Lukasik, A. Šeļa and M. Eder, Detecting direct speech in multilingual collection of 19th-century novels, in: *Proceedings of LT4HALA 2020 – 1st Workshop on Language Technologies for Historical and Ancient Languages*, R. Sprugnoli and M. Passarotti, eds, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 100–104. ISBN 979-10-95546-53-5.

[22] R. Cartuyvels, G. Spinks and M. Moens, Discrete and continuous representations and processing in deep learning: Looking forward, *AI Open* **2** (2021), 143–159. doi:10.1016/j.aiopen.2021.07.002.

[23] I. Chaturvedi, R. Satapathy, S. Cavallari and E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, *Pattern Recognition Letters* **125** (2019), 264–270. doi:10.1016/j.patrec.2019.04.024.

[24] S. Chaudhury, P. Sen, M. Ono, D. Kimura, M. Tatsubori and A. Munawar, Neuro-symbolic approaches for text-based policy learning, in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online, M.-F. Moens, X. Huang, L. Specia and S.W.-T. Yih, eds, Association for Computational Linguistics, Punta Cana, Dominican Republic 2021, pp. 3073–3078. doi:10.18653/v1/2021.emnlp-main.245.

[25] K. Chen, Q. Huang, H. Palangi, P. Smolensky, K.D. Forbus and J. Gao, Mapping natural-language problems to formal-language solutions using structured neural representations, in: *Proceedings of the 37th International Conference on Machine Learning, ICML'20*, JMLR.org, 2020.

[26] K. Chen, W. Xu, X. Cheng, Z. Xiaochuan, Y. Zhang, L. Song, T. Wang, Y. Qi and W. Chu, Question directed graph attention network for numerical reasoning over text, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He and Y. Liu, eds, Association for Computational Linguistics, 2020, pp. 6759–6768, online. doi:10.18653/v1/2020.emnlp-main.549.

[27] Q. Chen, A. Lamoreaux, X. Wang, G. Durrett, O. Bastani and I. Dillig, Web question answering with neurosymbolic program synthesis, in: *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*, S.N. Freund and E. Yahav, eds, PLDI 2021, Association for Computing Machinery, New York, NY, USA, 2021, pp. 328–343. ISBN 9781450383912. doi:10.1145/3453483.3454047.

[28] X. Chen, C. Liang, A.W. Yu, D. Song and D. Zhou, Compositional generalization via neural-symbolic stack machines, in: *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan and H. Lin, eds, Vol. 33, Curran Associates, Inc., 2020, pp. 1690–1701.

[29] Z. Chen, Q. Gao and L.S. Moss, NeuralLog: Natural language inference with joint neural and logical reasoning, in: *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics, *SEM 2021*, Online, August 5–6, 2021, V. Nastase and I. Vulic, eds, Association for Computational Linguistics, 2021, pp. 78–88. doi:10.18653/v1/2021.starsem-1.7.

[30] A.I. Cowen-Rivers, P. Minervini, T. Rocktaschel, M. Bosnjak, S. Riedel and J. Wang, Neural Variational Inference For Estimating Uncertainty in Knowledge Graph Embeddings, 2019, arXiv:1906.04985. doi:10.48550/ARXIV.1906.04985.

[31] Q. Cui, Y. Zhou and M. Zheng, Sememes-based framework for knowledge graph embedding with comprehensive-information, *Lecture Notes in Computer Science* **12816**(LNAI) (2021), 419–426. doi:10.1007/978-3-030-82147-0_34.

[32] R. Das, M. Zaheer, D. Thai, A. Godbole, E. Perez, J.Y. Lee, L. Tan, L. Polymenakos and A. McCallum, Case-based reasoning for natural language queries over knowledge bases, in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online, M. Moens, X. Huang, L. Specia and S.W. Yih, eds, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 9594–9611. doi:10.18653/v1/2021.emnlp-main.755.

[33] R. Davis, H.E. Shrobe and P. Szolovits, What is a knowledge representation?, *AI Magazine* **14**(1) (1993), 17–33. doi:10.1609/aimag.v14i1.1029.

[34] L. De Raedt, A. Kimmig and H. Toivonen, ProbLog: A probabilistic prolog and its application in link discovery, in: *Proceedings of the 20th International Joint Conference on Artifical Intelligence, IJCAI'07*, C. Bessiere, ed., Vol. 7, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2007, pp. 2468–2473.

[35] C. Dehua, Z. Keting and H. Jianrong, BDCN: Semantic embedding self-explanatory breast diagnostic capsules network, in: *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, S. Li, M. Sun, Y. Liu, H. Wu, K. Liu, W. Che, S. He and G. Rao, eds, Chinese Information Processing Society of China, Huhhot, China, 2021, pp. 1178–1189.

[36] D. Demeter and D. Downey, Just add functions: A neural-symbolic language model, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, AAAI Press, 2020, pp. 7634–7642. doi:10.1609/aaai.v34i05.6264.

[37] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long and Short Papers)*, J. Burstein, C. Doran and T. Solorio, eds, Vol. 1, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[38] M. Diligenti, M. Gori and C. Saccà, Semantic-based regularization for learning and inference, *Artificial Intelligence* **244** (2017), 143–165. doi:10.1016/j.artint.2015.08.011.

[39] H. Dong, J. Mao, T. Lin, C. Wang, L. Li and D. Zhou, Neural Logic Machines, 2019, arXiv:1904.11694. doi:10.48550/ARXIV.1904.11694.

[40] J. D'Souza, I.O. Mulang' and S. Auer, Team SVMrank: Leveraging feature-rich support vector machines for ranking explanations to elementary science questions, in: *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13), Association for Computational Linguistics*, Hong Kong, 2019, pp. 90–100. doi:10.18653/v1/D19-5312.

[41] C.R. Dyer, CS 540 Lecture Notes: Logic, University of Wisconsin – Madison.

[42] V. Embar, D. Sridhar, G. Farnadi and L. Getoor, Scalable Structure Learning for Probabilistic Soft Logic, 2018, arXiv:1807.00973. doi:10.48550/ARXIV.1807.00973.

[43] P. Engel, Reasoning and rationality, in: *Dictionary of Cognitive Science Neuroscience, Psychology, Artificial Intelligence, Linguistics, and Philosophy*, Taylor and Francis, 2003, pp. 315–316. doi:10.4324/9780203486030.

[44] F. Es-Sabery, A. Hair, J. Qadir, B. Sainz-De-Abajo, B. Garcia-Zapirain and I. Torre-DIez, Sentence-level classification using parallel fuzzy deep learning classifier, *IEEE Access* **9** (2021), 17943–17985. doi:10.1109/ACCESS.2021.3053917.

[45] R. Evans and E. Grefenstette, Learning explanatory rules from noisy data, *Journal of Artificial Intelligence Research* **61** (2018), 1–64. doi:10.1613/jair.5714.

[46] W. Farnsworth, *The Socratic Method: A Practitioner's Handbook*, David R. Godine Publisher Inc, 2021. ISBN 978-1-56792-685-9.

[47] L.B. Fazlic, A. Hallawa, A. Schmeink, A. Peine, L. Martin and G. Dartmann, A novel NLP-FUZZY system prototype for information extraction from medical guidelines, in: *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, M. Koricic, Z. Butkovic, K. Skala, Z. Car, M. Cicin-Sain, S. Babic, V. Sruk, D. Skvorc, S. Ribaric, S. Gros, B. Vrdoljak, M. Mauher, E. Tijan, P. Pale, D. Huljenic, T.G. Grbac and M. Janjic, eds, Opatija, Croatia, 2019, pp. 1025–1030. doi:10.23919/MIPRO.2019.8756929.

[48] D.A. Ferrucci, Introduction to "this is Watson", *IBM Journal of Research and Development* **56**(3.4) (2012), 1:1–1:15. doi:10.1147/JRD.2012.2184356.

[49] E. Gabrilovich, R. Guha, A. McCallum and K. Murphy, *Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches*, The AAAI Press, Palo Alto, California, 2015. ISBN 978-1-57735-707-0.

[50] A.D. Garcez, M. Gori, L.C. Lamb, L. Serafini, M. Spranger and S.N. Tran, Neural-Symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning, 2019, arXiv:1905.06088. doi:10.48550/ARXIV.1905.06088.

[51] A.D. Garcez and L.C. Lamb, Neurosymbolic AI: The 3rd Wave, 2020, arXiv:2012.05876. doi:10.48550/ARXIV.2012.05876.

[52] A.S. Garcez, L.C. Lamb and D.M. Gabbay, *Neural-Symbolic Cognitive Reasoning, Cognitive Technologies*, Springer, 2009. ISBN 978-3-540-73245-7. doi:10.1007/978-3-540-73246-4.

[53] A.S.D. Garcez, K. Broda, D.M. Gabbay et al., *Neural-Symbolic Learning Systems: Foundations and Applications*, Springer Science & Business Media, 2002. doi:10.1007/978-1-4471-0211-3.

[54] A.S.D. Garcez and D.M. Gabbay, Fibring neural networks, in: *Proceedings of 19th National Conference on Artificial Intelligence – AAAI-2004*, A.G. Cohn, ed., AAAI Press, 2004, pp. 342–347. doi:10.5555/1597148.

[55] A. Gatt and E. Krahmer, Survey of the state of the art in natural language generation: Core tasks, applications and evaluation, *Journal of Artificial Intelligence Research* **61** (2018), 65–170. doi:10.1613/jair.5477.

[56] J. Gong, H. Ma, Z. Teng, Q. Teng, H. Zhang, L. Du, S. Chen, M.Z.A. Bhuiyan, J. Li and M. Liu, Hierarchical graph transformer-based deep learning model for large-scale multi-label text classification, *IEEE Access* **8** (2020), 30885–30896. doi:10.1109/ACCESS.2020.2972751.

[57] A.S. Gordon and J.R. Hobbs, *A Formal Theory of Commonsense Psychology: How People Think People Think*, Cambridge University Press, 2017. doi:10.1017/9781316584705.

[58] L. Graziani, S. Melacci and M. Gori, Jointly learning to detect emotions and predict Facebook reactions, in: *Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series*, I.V. Tetko, V. Kůrková, P. Karpov and F. Theis, eds, Springer International Publishing, Cham, 2019, pp. 185–197. ISBN 978-3-030-30490-4. doi:10.1007/978-3-030-30490-4_16.

[59] Y. Gu, J.Z. Pan, G. Cheng, H. Paulheim and G. Stoilos, Local ABox consistency prediction with transparent TBoxes using gated graph neural networks, in: *Proceedings of the 2019 International Workshop on Neural-Symbolic Learning and Reasoning*, D. Doran, A. d'Avila Garcez and F. Lecue, eds, 2019, pp. 48–53.

[60] K. Gupta, T. Ghosal and A. Ekbal, A neuro-symbolic approach for question answering on research articles, in: *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, K. Hu, J. Kim, C. Zong and E. Chersoni, eds, Association for Computational Lingustics, Shanghai, China, 2021, pp. 40–49.

[61] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman and N.A. Smith, Annotation artifacts in natural language inference data, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, M.A. Walker, H. Ji and A. Stent, eds, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 107–112. doi:10.18653/v1/N18-2017.

[62] W.L. Hamilton, R. Ying and J. Leskovec, *Representation Learning on Graphs: Methods and Applications*, 2017, arXiv:1709.05584. doi:10.48550/ARXIV.1709.05584.

[63] B. Hammer and P. Hitzler (eds), *Perspectives of Neural-Symbolic Integration*, Vol. 77, Springer, 2007. ISBN 978-3-540-73953-1.

[64] F. Harder and T.R. Besold, *Cognitive Systems Research* **47** (2018), 42–67. doi:10.1016/j.cogsys.2017.07.004.

[65] P. Hitzler, F. Bianchi, M. Ebrahimi and M.K. Sarker, Neural-symbolic integration and the semantic web, *Semantic Web* **11**(1) (2020), 3–11. doi:10.3233/SW-190368.

[66] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Computation* **9**(8) (1997), 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[67] H. Honda and M. Hagiwara, Question answering systems with deep learning-based symbolic processing, *IEEE Access* **7** (2019), 152368–152378. doi:10.1109/ACCESS.2019.2948081.

[68] D. Hu, L. Wei and X. Huai, DialogueCRN: Contextual reasoning networks for emotion recognition in conversations, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li and R. Navigli, eds, Association for Computational Linguistics, 2021, pp. 7042–7052, Virtual Event. doi:10.18653/v1/2021.acl-long.547.

[69] Q. Huang, L. Deng, D. Wu, C. Liu and X. He, Attentive tensor product learning, *Proceedings of the AAAI Conference on Artificial Intelligence* **33** (2019), 1344–1351. doi:10.1609/aaai.v33i01.33011344.

[70] S. Huo, T. Ma, J. Chen, M. Chang, L. Wu and M. Witbrock, Graph enhanced cross-domain text-to-SQL generation, in: *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, D. Ustalov, S. Somasundaran, P. Jansen, G. Glavaš, M. Riedl, M. Surdeanu and M. Vazirgiannis, eds, Association for Computational Linguistics, 2019, pp. 159–163. doi:10.18653/v1/D19-5319.

[71] A. Hussain and E. Cambria, Semi-supervised learning for big social data analysis, *Neurocomputing* **275** (2018), 1662–1673. doi:10.1016/j.neucom.2017.10.010.

[72] R. Jia and P. Liang, Adversarial examples for evaluating reading comprehension systems, in: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa and S. Riedel, eds, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2021–2031. doi:10.18653/v1/d17-1215.

[73] H. Jiang, S. Gurajada, Q. Lu, S. Neelam, L. Popa, P. Sen, Y. Li and A. Gray, LNN-EL: A neuro-symbolic approach to short-text entity linking, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online, C. Zong, F. Xia, W. Li and R. Navigli, eds, Association for Computational Linguistics, 2021, pp. 775–787. doi:10.18653/v1/2021.acl-long.64.

[74] H. Jiang, P. He, W. Chen, X. Liu, J. Gao and T. Zhao, SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter and J. Tetreault, eds, Association for Computational Linguistics, 2020, pp. 2177–2190, online. doi:10.18653/v1/2020.acl-main.197.

[75] J. Jiang, H. Wang, J. Xie, X. Guo, Y. Guan and Q. Yu, Medical knowledge embedding based on recursive neural network for multi-disease diagnosis, *Artificial Intelligence in Medicine* **103** (2020), 101772. doi:10.1016/j.artmed.2019.101772.

[76] M.F. Joanisse and J.L. McClelland, Connectionist perspectives on language learning, representation and processing, *Wiley Interdisciplinary Reviews: Cognitive Science* **6**(3) (2015), 235–247.

[77] D. Kahneman, *Thinking, Fast and Slow*, Farrar, Straus and Giroux, New York, 2011, 0374275637. ISBN 9780374275631.

[78] D. Kahneman, O. Sibony and C.R. Sunstein, *Noise: A Flaw in Human Judgment*, HarperCollins Publishers, Limited, 2021. ISBN 978-0-00-830900-8.

[79] M. Kang and N.J. Jameson, Machine learning: Fundamentals, in: *Prognostics and Health Management of Electronics*, M.G. Pecht and M. Kang, eds, John Wiley & Sons, Ltd, 2018, pp. 85–109, Chapter 4. ISBN 9781119515326. doi:10.1002/9781119515326.ch4.

[80] H. Kautz, The Third AI Summer, AAAI Robert S. Engelmore Memorial Lecture, Thirty-fourth AAAI Conference on Artificial Intelligence, New York, NY, https://henrykautz.com/talks/index.html.

[81] T.N. Kipf and M. Welling, Semi-supervised classification with graph convolutional networks, 2017, arXiv:1609.02907. doi:10.48550/arXiv.1609.02907.

[82] B. Kitchenham, *Procedures for Performing Systematic Reviews*, Vol. 33, Keele University, Keele, UK, 2004, pp. 1–26.

[83] K. Kogkalidis, M. Moortgat and R. Moot, Neural proof nets, in: *Proceedings of the 24th Conference on Computational Natural Language Learning, CoNLL 2020*, Online, November 19–20, 2020, R. Fernández and T. Linzen, eds, Association for Computational Linguistics, 2020, pp. 26–40. doi:10.18653/v1/2020.conll-1.3.

[84] D. Koller, N. Friedman, S. Džeroski, C. Sutton, A. McCallum, A. Pfeffer, P. Abbeel, M.-F. Wong, C. Meek, J. Neville et al., *Introduction to Statistical Relational Learning*, MIT Press, 2007.

[85] P. Kouris, G. Alexandridis and A. Stafylopatis, Abstractive text summarization: Enhancing sequence-to-sequence models using word sense disambiguation and semantic content generalization, *Computational Linguistics* **47**(4) (2021), 813–859. doi:10.1162/coli_a_00417.

[86] G. Lakoff, Linguistics and natural logic, *Synthese* **22**(1) (1970), 151–271. doi:10.1007/BF00413602.

[87] L.C. Lamb, A.D. Garcez, M. Gori, M.O.R. Prates, P.H.C. Avelar and M.Y. Vardi, Graph neural networks meet neural-symbolic computing: A survey and perspective, in: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, C. Bessiere, ed., International Joint Conferences on Artificial Intelligence Organization, 2020, pp. 4877–4884. ISBN 978-0-9992411-6-5. doi:10.24963/ijcai.2020/679.

[88] G. Lample and F. Charton, Deep learning for symbolic mathematics, in: *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net, 2020.

[89] J. Langton and K. Srihasam, Applied medical code mapping with character-based deep learning models and word-based logic, in: *Proceedings of the 1st and 2nd Workshops on Natural Logic Meets Machine Learning (NALOMA)*, A.-L. Kalouli and L.S. Moss, eds, Association for Computational Linguistics, Groningen, the Netherlands, 2021, pp. 7–11, online.

[90] J. Lehman and R. Miikkulainen, Neuroevolution, *Scholarpedia* **8**(6) (2013), 30977. doi:10.4249/scholarpedia.30977.

[91] H. Lemos, P. Avelar, M. Prates, A. Garcez and L. Lamb, in: *Neural-Symbolic Relational Reasoning on Graph Models: Effective Link Inference and Computation from Knowledge Bases*, Lecture Notes in Computer Science 12396 LNCS, 2020, pp. 647–659. doi:10.1007/978-3-030-61609-0_51.

[92] H.J. Levesque, Knowledge representation and reasoning, *Annual Review of Computer Science* **1**(1) (1986), 255–287. doi:10.1146/annurev.cs.01.060186.001351.

[93] T. Li and V. Srikumar, Augmenting neural networks with first-order logic, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D.R. Traum and L. Màrquez, eds, Association for Computational Linguistics, Florence, Italy, 2019, pp. 292–302. doi:10.18653/v1/P19-1028.

[94] R. Lima, B. Espinasse and F.L.G. de Freitas, The impact of semantic linguistic features in relation extraction: A logical relational learning approach, in: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, R. Mitkov and G. Angelova, eds, INCOMA Ltd., Varna, Bulgaria, 2019, pp. 648–654. doi:10.26615/978-954-452-056-4_076.

[95] W. Liu, J. Tang, X. Liang and Q. Cai, Heterogeneous graph reasoning for knowledge-grounded medical dialogue system, *Neurocomputing* **442** (2021), 260–268. doi:10.1016/j.neucom.2021.02.021.

[96] Z. Liu, Z. Wang, Y. Lin and H. Li, A Neural-Symbolic Approach to Natural Language Understanding, 2022, arXiv:2203.10557. doi:10.48550/arXiv.2203.10557.

[97] B. MacCartney, Understanding Natural Language Understanding, ACM SIGAI Bay Area Chapter Inaugural Meeting, San Mateo, CA, https://www.youtube.com/watch?v=vcPd0V4VSNU.

[98] B. MacCartney and C.D. Manning, Natural logic for textual inference, in: *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, S. Sekine, K. Inui, I. Dagan, B. Dolan, D. Giampiccolo and B. Magnini, eds, Association for Computational Linguistics, Prague, 2007, pp. 193–200. doi:10.3115/1654536.1654575.

[99] B. MacCartney and C.D. Manning, An extended model of natural logic, in: *Proceedings of the Eight International Conference on Computational Semantics*, H. Bunt, ed., Association for Computational Linguistics, Tilburg, The Netherlands, 2009, pp. 140–156.

[100] P. Manda, S. SayedAhmed and S.D. Mohanty, Automated ontology-based annotation of scientific literature using deep learning, in: *Proceedings of the International Workshop on Semantic Big Data*, S. Groppe, L. Gruenwald and V. Presutti, eds, Vol. SBD'20, Association for Computing Machinery, New York, NY, USA, 2020. ISBN 9781450379748. doi:10.1145/3391274.3393636.

[101] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester and L. De Raedt, DeepProbLog: Neural probabilistic logic programming, in: *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds, Vol. 31, Curran Associates, Inc., 2018.

[102] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard and D. McClosky, The Stanford CoreNLP natural language processing toolkit, in: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, K. Bontcheva and J. Zhu, eds, Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 55–60. doi:10.3115/v1/P14-5010.

[103] J. Mao, C. Gan, P. Kohli, J.B. Tenenbaum and J. Wu, The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision, in: *7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA, USA, May 6–9, 2019, OpenReview.net, 2019.

[104] G. Marcus, Deep Learning: A Critical Appraisal, 2018, arXiv:1801.00631. doi:10.48550/ARXIV.1801.00631.

[105] G. Marcus and E. Davis, GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about |, *MIT Technology Review.*, https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/.

[106] S. McCall, Review of selected works, *Synthese* **26**(1) (1973), 165–171.

[107] J.P. McCrae, E. Rudnicka and F. Bond, 2021, English WordNet: A new open-source wordnet for English.

[108] R. Miikkulainen, Neuroevolution, in: *Encyclopedia of Machine Learning*, Springer, New York, 2010.

[109] S. Muggleton, Inductive logic programming, *New Generation Computing* **8**(4) (1991), 295–318. doi:10.1007/BF03037089.

[110] M.L. Pacheco and D. Goldwasser, Modeling content and context with deep relational learning, *Transactions of the Association for Computational Linguistics* **9** (2021), 100–119. doi:10.1162/tacl_a_00357.

[111] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J.M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting and D. Moher, The PRISMA 2020 statement: An updated guideline for reporting systematic reviews, *Systematic Reviews* **10**(1) (2021), 89. doi:10.1186/s13643-021-01626-4.

[112] G. Paré, M.-C. Trudel, M. Jaana and S. Kitsiou, Synthesizing information systems knowledge: A typology of literature reviews, *Information & Management* **52**(2) (2015), 183–199. doi:10.1016/j.im.2014.08.008.

[113] J. Pearl, *Theoretical Impediments to Machine Learning with Seven Sparks from the Causal Revolution*, 2018, arXiv:1801.04016. doi:10.48550/ARXIV.1801.04016.

[114] C. Pinhanez, P. Cavalin, V.H. Alves Ribeiro, A. Appel, H. Candello, J. Nogima, M. Pichiliani, M. Guerra, M. de Bayser, G. Malfatti and H. Ferreira, Using meta-knowledge mined from identifiers to improve intent recognition in conversational systems, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li and R. Navigli, eds, Association for Computational Linguistics, 2021, pp. 7014–7027, online. doi:10.18653/v1/2021.acl-long.545.

[115] S. Pinker, Words and rules, *Lingua* **106**(1–4) (1998), 219–242. doi:10.1016/S0024-3841(98)00035-7.

[116] J. Qin, X. Liang, Y. Hong, J. Tang and L. Lin, Neural-symbolic solver for math word problems with auxiliary tasks, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li and R. Navigli, eds, Association for Computational Linguistics, 2021, pp. 5870–5881, online. doi:10.18653/v1/2021.acl-long.456.

[117] Reasoning, Encyclopædia Britannica, inc. https://www.britannica.com/technology/artificial-intelligence/Reasoning.

[118] M. Richardson and P. Domingos, Markov logic networks, *Machine Learning* **62**(1) (2006), 107–136. doi:10.1007/s10994-006-5833-1.

[119] R. Riegel, A. Gray, F. Luus, N. Khan, N. Makondo, I.Y. Akhalwaya, H. Qian, R. Fagin, F. Barahona, U. Sharma, S. Ikbal, H. Karanam, S. Neelam, A. Likhyani and S. Srivastava, Logical Neural Networks, 2020, arXiv:2006.13155. doi:10.48550/ARXIV.2006.13155.

[120] T. Rocktäschel and S. Riedel, Learning knowledge base inference with neural theorem provers, in: *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, J. Pujara, T. Rocktaschel, D. Chen and S. Singh, eds, Association for Computational Linguistics, San Diego, CA, 2016, pp. 45–50. doi:10.18653/v1/W16-13.

[121] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova and C. Zhong, Interpretable machine learning: Fundamental principles and 10 grand challenges, *Statistics Surveys* **16** (2022), 1–85. doi:10.1214/21-SS133.

[122] H. Santos, M. Kejriwal, A.M. Mulvehill, G. Forbush and D.L. McGuinness, An experimental study measuring human annotator categorization agreement on commonsense sentences, *Experimental Results* **2** (2021), e19. doi:10.1017/exp.2021.9.

[123] M.K. Sarker, L. Zhou, A. Eberhart and P. Hitzler, *Neuro-Symbolic Artificial Intelligence: Current Trends*, 2021, arXiv:2105.05330. doi:10.48550/ARXIV.2105.05330.

[124] E. Saveleva, V. Petukhova, M. Mosbach and D. Klakow, Graph-based argument quality assessment, in: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, R. Mitkov, G. Angelova and K. Bontcheva, eds, INCOMA Ltd. Shoumen, BULGARIA, Held Online, 2021, pp. 1268–1280.

[125] F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner and G. Monfardini, The graph neural network model, *IEEE transactions on neural networks* **20**(1) (2008), 61–80. doi:10.1109/TNN.2008.2005605.
[126] M. Schlichtkrull, T.N. Kipf, P. Bloem, R.V.D. Berg, I. Titov and M. Welling, Modeling relational data with graph convolutional networks, in: *European Semantic Web Conference*, A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai and M. Alam, eds, Vol. 10843, Springer International Publishing, 2018, pp. 593–607. doi:10.1007/978-3-319-93417-4_38.
[127] C. Schon, S. Siebert and F. Stolzenburg, The CoRg project: Cognitive reasoning, *Künstliche Intell.* **33**(3) (2019), 293–299. doi:10.1007/s13218-019-00601-5.
[128] P. Sen, M. Danilevsky, Y. Li, S. Brahma, M. Boehm, L. Chiticariu and R. Krishnamurthy, Learning explainable linguistic expressions with neural inductive logic programming for sentence classification, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He and Y. Liu, eds, Association for Computational Linguistics, 2020, pp. 4211–4221, online. doi:10.18653/v1/2020.emnlp-main.345.
[129] L. Serafini and A.S. d'Avila Garcez, Learning and reasoning with logic tensor networks, in: *AI*IA 2016 Advances in Artificial Intelligence*, G. Adorni, S. Cagnoni, M. Gori and M. Maratea, eds, Lecture Notes in Computer Science, Springer International Publishing, 2016, pp. 334–348. ISBN 978-3-319-49130-1. doi:10.1007/978-3-319-49130-1_25.
[130] L. Serafini and A.D. Garcez, Logic Tensor Networks: Deep Learning and Logical Reasoning from Data and Knowledge, 2016, arXiv:1606.04422. doi:10.48550/ARXIV.1606.04422.
[131] O. Sharir, B. Peleg and Y. Shoham, The Cost of Training NLP Models: A Concise Overview, ArXiv, 2020. doi:10.48550/arXiv.2004.08900.
[132] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu and P. Cui, Towards Out-Of-Distribution Generalization: A Survey, 2021, arXiv:2108.13624. doi:10.48550/ARXIV.2108.13624.
[133] J. Shi, X. Ding, L. Du, T. Liu and B. Qin, Neural natural logic inference for interpretable question answering, in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M. Moens, X. Huang, L. Specia and S.W. Yih, eds, Association for Computational Linguistics, 2021, pp. 3673–3684. doi:10.18653/v1/2021.emnlp-main.298.
[134] B. Škrlj, M. Martinc, N. Lavrač and S. Pollak, autoBOT: Evolving neuro-symbolic representations for explainable low resource text classification, *Machine Learning* **110**(5) (2021), 989–1028. doi:10.1007/s10994-021-05968-x.
[135] P. Smolensky, Tensor product variable binding and the representation of symbolic structures in connectionist systems, *Artificial Intelligence* **46**(1–2) (1990), 159–216. doi:10.1016/0004-3702(90)90007-M.
[136] R. Socher, D. Chen, C.D. Manning and A. Ng, Reasoning with neural tensor networks for knowledge base completion, in: *Advances in Neural Information Processing Systems*, C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani and K.Q. Weinberger, eds, Vol. 26, Curran Associates, Inc., 2013.
[137] J.F. Sowa, *Principles of Semantic Networks*, Morgan Kaufmann, 1991. ISBN 978-1-4832-0771-1. doi:10.1016/C2013-0-08297-7.
[138] R.D. Sriram, Analogical and case-based reasoning, in: *Intelligent Systems for Engineering: A Knowledge-Based Approach*, Springer London, London, 1997, pp. 285–334. ISBN 978-1-4471-0631-9. doi:10.1007/978-1-4471-0631-9_6.
[139] C. Strasser and G.A. Antonelli, Non-monotonic logic, in: *The Stanford Encyclopedia of Philosophy, Summer 2019 edn*, 2019, Research Lab, Stanford University.
[140] A. Sutherland, S. Magg and S. Wermter, Leveraging recursive processing for neural-symbolic affect-target associations, in: *International Joint Conference on Neural Networks, IJCNN, 2019*, Budapest, Hungary, July 14–19, 2019, IEEE, 2019, pp. 1–6. doi:10.1109/IJCNN.2019.8851875.
[141] A.A.N. Tato, R. Nkambou and A. Dufresne, Hybrid deep neural networks to predict socio-moral reasoning skills, in: *Proceedings of the 12th International Conference on Educational Data Mining*, C.F. Lynch, A. Merceron, M. Desmarais and R. Nkambou, eds, International Educational Data Mining Society (IEDMS), 2019, pp. 623–626.
[142] G.G. Towell and J.W. Shavlik, Knowledge-based artificial neural networks, *Artificial intelligence* **70**(1–2) (1994), 119–165.
[143] I.L. Travis, Knowledge representation in artificial intelligence, in: *Clinic on Library Applications of Data Processing*, Vol. 27, 1990, p. 1990.
[144] E. Tsamoura, T. Hospedales and L. Michael, Neural-symbolic integration: A compositional perspective, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 2021, pp. 5051–5060.
[145] L.G. Valiant, Three problems in computer science, *Journal of the ACM* **50**(1) (2003), 96–99. doi:10.1145/602382.602410.
[146] F. Van Harmelen and A.T. Teije, A boxology of design patterns for hybrid learning and reasoning systems, *Journal of Web Engineering* **18** (2019), 97–124. doi:10.13052/jwe1540-9589.18133.
[147] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.U. Kaiser and I. Polosukhin, Attention is all you need, in: *Advances in Neural Information Processing Systems*, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds, Vol. 30, Curran Associates, Inc., 2017, pp. 5998–6008.
[148] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò and Y. Bengio, *Graph Attention Networks, CoRR* (2017), arXiv:1710.10903.
[149] P. Verga, H. Sun, L. Baldini Soares and W. Cohen, Adaptable and interpretable neural MemoryOver symbolic knowledge, in: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty and Y. Zhou, eds, Association for Computational Linguistics, 2021, pp. 3678–3691. doi:10.18653/v1/2021.naacl-main.288.
[150] G. Vilone and L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, *Information Fusion* **76** (2021), 89–106. doi:10.1016/j.inffus.2021.05.009.
[151] O. Vinyals, M. Fortunato and N. Jaitly, Pointer networks, in: *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama and R. Garnett, eds, Vol. 28, Curran Associates, Inc., 2015.

[152] L. von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, M. Walczak, J. Pfrommer, A. Pick, R. Ramamurthy, J. Garcke, C. Bauckhage and J. Schuecker, Informed Machine Learning – A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems, *IEEE Transactions on Knowledge and Data Engineering* (2021), 1–1. doi:10.1109/TKDE.2021.3079836.

[153] W. Wang and S.J. Pan, Variational deep logic network for joint inference of entities and relations, *Computational Linguistics* **47**(4) (2021), 775–812. doi:10.1162/coli_a_00415.

[154] Y. Wang, Q. Yao, J.T. Kwok and L.M. Ni, Generalizing from a few examples: A survey on few-shot learning, *ACM computing surveys (csur)* **53**(3) (2020), 1–34.

[155] J. Weizenbaum, ELIZA – a computer program for the study of natural language communication between man and machine, *Communications of the ACM* **9**(1) (1966), 36–45. doi:10.1145/365153.365168.

[156] C. Wendelken and L. Shastri, Multiple instantiation and rule mediation in SHRUTI, *Connection Science* **16**(3) (2004), 211–217. doi:10.1080/09540090412331311932.

[157] M. Wu, W. Wang and S.J. Pan, Deep weighted MaxSAT for aspect-based opinion extraction, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, B. Webber, T. Cohn, Y. He and Y. Liu, eds, Association for Computational Linguistics, 2020, pp. 5618–5628. doi:10.18653/v1/2020.emnlp-main.453.

[158] C. Xu and R. Li, Relation embedding with dihedral group in knowledge graph, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D.R. Traum and L. Màrquez, eds, Association for Computational Linguistics, Florence, Italy, 2019, pp. 263–272. doi:10.18653/v1/P19-1026.

[159] L. Yabloko, ETHAN at SemEval-2020 task 5: Modelling causal reasoning in language using neuro-symbolic cloud computing, in: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May and E. Shutova, eds, International Committee for Computational Linguistics, Barcelona, 2020, pp. 645–652, online. doi:10.18653/v1/2020.semeval-1.83.

[160] F. Yang, Z. Yang and W.W. Cohen, Differentiable learning of logical rules for knowledge base reasoning, *Advances in neural information processing systems* **30** (2017).

[161] L. Yao, C. Mao and Y. Luo, Graph convolutional networks for text classification, in: *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'19/IAAI'19/EAAI'19*, Vol. 33, AAAI Press, 2019, pp. 7370–7377. ISBN 978-1-57735-809-1. doi:10.1609/aaai.v33i01.33017370.

[162] Y. Yao, J. Xu, J. Shi and B. Xu, Learning to activate logic rules for textual reasoning, *Neural Networks* **106** (2018), 42–49. doi:10.1016/j.neunet.2018.06.012.

[163] D. Yu, B. Yang, D. Liu and H. Wang, A Survey on Neural-symbolic Systems, 2021, arXiv:2111.08164. doi:10.48550/ARXIV.2111.08164.

[164] J. Zhang, B. Chen, L. Zhang, X. Ke and H. Ding, Neural, symbolic and neural-symbolic reasoning on knowledge graphs, *AI Open* **2** (2021), 14–35. doi:10.1016/j.aiopen.2021.03.001.

[165] Q. Zhang, L. Wang, S. Yu, S. Wang, Y. Wang, J. Jiang and E.-P. Lim, NOAHQA: Numerical reasoning with interpretable graph question answering dataset, in: *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia and S.W.-T. Yih, eds, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 4147–4161. doi:10.18653/v1/2021.findings-emnlp.350.

[166] B. Zhou, K. Richardson, Q. Ning, T. Khot, A. Sabharwal and D. Roth, Temporal reasoning on implicit events from distant supervision, in: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty and Y. Zhou, eds, Association for Computational Linguistics, 2021, pp. 1361–1371, online. doi:10.18653/v1/2021.naacl-main.107.

[167] M. Zhou, D. Ji and F. Li, Relation extraction in dialogues: A deep learning model based on the generality and specialty of dialogue text, *IEEE/ACM Transactions on Audio Speech and Language Processing* **29** (2021), 2015–2026. doi:10.1109/TASLP.2021.3082295.

[168] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong and Q. He, A comprehensive survey on transfer learning, *Proceedings of the IEEE* **109**(1) (2021), 43–76. doi:10.1109/JPROC.2020.3004555.