Revisit Systematic Generalization via Meaningful Learning

*Ning Shi♠♡ *Boxin Wang♣ Wei Wang♡ Xiangyu Liu♡ †Zhouhan Lin★
♠Alberta Machine Intelligence Institute, Dept. of Computing Science, University of Alberta
♣University of Illinois at Urbana-Champaign
♡Alibaba Group ★Shanghai Jiao Tong University
ning.shi@ualberta.ca, boxinw2@illinois.edu
{luyang.ww,eason.lxy}@alibaba-inc.com, lin.zhouhan@gmail.com

Abstract

Humans can systematically generalize to novel compositions of existing concepts. Recent studies argue that neural networks appear inherently ineffective in such cognitive capacity, leading to a pessimistic view and a lack of attention to optimistic results. We revisit this controversial topic from the perspective of meaningful learning, an exceptional capability of humans to learn novel concepts by connecting them with known ones. We reassess the compositional skills of sequence-to-sequence models conditioned on the semantic links between new and old concepts. Our observations suggest that models can successfully one-shot generalize to novel concepts and compositions through semantic linking, either inductively or deductively. We demonstrate that prior knowledge plays a key role as well. In addition to synthetic tests, we further conduct proof-of-concept experiments in machine translation and semantic parsing, showing the benefits of meaningful learning in applications. We hope our positive findings will encourage excavating modern neural networks' potential in systematic generalization through more advanced learning schemes.

1 Introduction

As a crucial characteristic of human cognition, systematic generalization reflects people's talents to learn infinite combinations of finite concepts (Chomsky, 1956; Montague et al., 1970). Whether connectionist networks can express language and thoughts systematically has been controversial for many years (Fodor and Pylyshyn, 1988; Hadley, 1994; Marcus, 1998; Fodor and Lepore, 2002; Brakel and Frank, 2009; Frank et al., 2009; Marcus, 2018). To date, the systematic compositionality in neural networks remains an appealing research topic. Evidence on multiple explicitly proposed language-based generalization challenges suggests

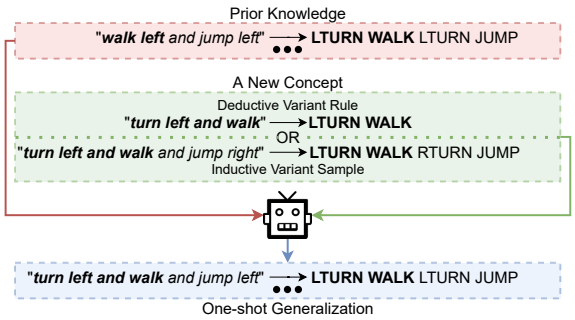

Figure 1: An example of the one-shot compositional generalization from the old concept "***walk left***" to the new one "***turn left and walk***" in SCAN. The model is able to generalize from the command "***walk left*** *and jump left*" to "***turn left and walk*** *and jump left*" through the semantic relationship between the old and new concepts because they refer to the same action "**LTURN WALK**". Such semantic linking can be established by either an inductive sample or a deductive rule.

that models lack such cognitive capacity (Bastings et al., 2018; Loula et al., 2018; Sinha et al., 2019; Keysers et al., 2020; Hupkes et al., 2020; Kim and Linzen, 2020; Li et al., 2021). Tremendous efforts are made to tackle these challenges through architectural modifications (Li et al., 2019; Gordon et al., 2020; Oren et al., 2020; Akyurek and Andreas, 2021; Chaabouni et al., 2021), meta-learning (Lake, 2019; Conklin et al., 2021), grammar (Kim, 2021; Shaw et al., 2021), neuro-symbolic models (Chen et al., 2020; Liu et al., 2020; Nye et al., 2020), data augmentation (Andreas, 2020; Akyürek et al., 2021; Auersperger and Pecina, 2021; Jiang and Bansal, 2021; Patel et al., 2022), and loss design (Yin et al., 2021). Despite their astounding accomplishments, standard sequence-to-sequence (seq2seq) models (Sutskever et al., 2014) appear to have relatively weak inductive biases, failing to capture underlying hierarchical structure.

In contrast, the successful one-shot generalization in the turn-left experiment on the **S**implified **C**omm**AI** **N**avigation (SCAN) task reveals the potential of seq2seq recurrent networks in controlled

* Work was done at Alibaba Group.
† Zhouhan Lin is the corresponding author.

Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 62 - 79

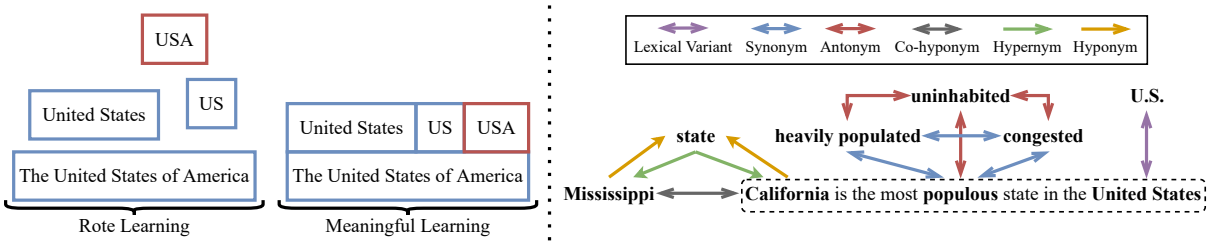

Figure 2: Adapted examples from Geography. In the left one, intuitively, knowing how the new concept (e.g., "*USA*") relates to the other existing ones (e.g., "*US*") can boost the learning and memory of this knowledge as a whole. In the right one, bidirectional arrows denote symmetric relations. "*Mississippi*" and "*California*" are two specific states, and thus both are hyponyms of "*state*". In turn, "*state*" is a hypernym of them. Due to a common hypernym, "*Mississippi*" and "*California*" become a co-hyponym for each other. {"*heavily populated*", "*congested*", "*populus*"} is a group of synonyms as sharing similar semantics. Finally, "*U.S.*", as a kind of abbreviation, is a lexical variant of "*United States*".

environments (Lake and Baroni, 2018). Although models are only exposed to the primitive command before, they are able to understand most composed commands of "*turn left*". One assumption is that models study new commands with a primitive from other action sequences containing the basic action it denotes. However, there is still a missing formal exploration to answer the question raised by Lake and Baroni (2018) on page 8 that "*what are, precisely, the generalization mechanisms that subtend the networks' success in these experiments*".

In this work, as a response to the call, we question whether neural networks are indeed deficient or just conventional learning protocols unable to exploit their full potential (Csordás et al., 2021; Dankers et al., 2022b). We revisit the systematic generalization of seq2seq models from a *meaningful learning* perspective (Ausubel, 1963; Okebukola and Jegede, 1988; Mayer, 2002). Given the idea that humans are used to memorizing concepts in a relational manner, we hypothesize that the success of the turn-left experiment results from the semantic relationships between old concepts and new ones. For example, in Figure 1, a model can understand the meaning of "***turn left and walk*** *and jump left*" from "***walk left*** *and jump left*" via the semantic link between two concepts (in bold) since both denote to the same action "LTURN WALK".

To validate our hypothesis, we reproduce the one-shot compositional generalization by *semantic linking* that exposes semantic relationships through either *inductive learning* or *deductive learning* (Hammerly, 1975; Shaffer, 1989; Thornbury, 1999). On the one hand, by introducing new concepts sharing the same context, we hope the model can capture the underlying semantic connections inductively. On the other hand, by involving a rule-like concept dictionary without specific context information, we hope the model can utilize the general cross-lingual supervised signals as anchor points so as to launch the semantic linking deductively.

In experiments, we treat concepts in the initial data set as primitives and generate variant samples and rules accordingly. Next, we mix them up and construct a seq2seq task after a random split. We repeatedly train and evaluate models but slowly decrease the number of times they see each variant until one-shot learning. We observe there is hardly a performance drop in SCAN for three representative model structures. This evidences that, with semantic linking, even canonical neural networks can generalize systematically to new concepts and compositions. Such observation holds consistently across two more semantic parsing (SP) datasets. The followed sensitivity analysis shows that prior knowledge also takes essential parts. Lastly, as a proof-of-concept, we demonstrate how meaningful learning already benefits models in standard machine translation (MT) and SP. Overall, our contributions[1] are as follows:

- We revisit systematic generalization from a meaningful learning perspective by either inductive or deductive semantic linking.
- We find that modern seq2seq models can generalize to new concepts and compositions after semantic linking, which empirically answers the question by Lake and Baroni (2018).
- We show in the sensitivity analysis that both semantic linking and prior knowledge play a key role, in line with meaningful learning theory.
- We extend to standard MT and SP and demonstrate how meaningful learning already benefits models in solving realistic problems.

[1]Code and data are publicly available at GitHub.

2 Meaningful Learning

In educational psychology, meaningful learning refers to learning new concepts by relating them to old ones (Ausubel, 1963; Mayer, 2002). In Figure 2, intuitively, the utilization of meaningful learning can encourage learners to understand information continuously built on concepts the learners already understand (Okebukola and Jegede, 1988). Following this, we intend to examine models' systematic compositionality by exploring semantic linking that establishes semantic relations between primitives (old concepts) and their variants (new concepts). We propose to spoon-feed semantic knowledge to models for semantic linking in two ways, that is, inductive learning and deductive learning (Hammerly, 1975; Shaffer, 1989; Thornbury, 1999). In this section, we discuss the process of semantic linking and take "*jump*" from SCAN as an example primitive to illustrate the learning scheme.

2.1 Semantic links

We focus on three semantic relationships, namely, *lexical variant*, *co-hyponym*, and *synonym*. Lexical Variant refers to an alternative expression form for the same concept. Co-hyponym is a linguistic term to designate a semantic relation between two group members belonging to the same broader class, where each member is a hyponym and the class is a hypernym (Lyons and John, 1995). Synonym stands for a word, morpheme, or phrase that shares exactly or nearly the same semantics with another one. We provide an example in Figure 2 and a detailed description in Appendix A.

2.2 Inductive learning

Inductive learning is a bottom-up approach from the more specific to the more general. In grammar teaching, inductive learning is a rule-discovery approach starting with the presentation of specific examples from which a general rule can be inferred (Thornbury, 1999). In semantic linking, we propose to introduce variant samples sharing the same context with their primitives during training. The assumption is that models can observe the interchange of primitives and their variants surrounded by the same context in the hope of coming up with a general hypothesis that there is a semantic linking between primitives and their variants (Harris, 1954). To test the generalization, we design a prompt "*[concept] twice*" from a primitive sample "*jump twice*". After that, we fill in the concept slot with "*jump_0*" and generate the variant sample "*jump_0 twice*". There is no change from the target side. Finally, by training models on the generated variant sample in combination with prior knowledge (all the other primitive samples), we aim to establish the semantic relationship between "*jump*" and "*jump_0*" inductively.

2.3 Deductive learning

Deductive Learning, the opposite of inductive learning, is a top-down approach from the more general to the more specific. As a rule-driven approach, teaching in a deductive manner often begins with presenting a general rule followed by specific examples in practice where the rule is applied (Thornbury, 1999). To align with this definition, we intend to do semantic linking deductively by combining a bilingual dictionary that maps primitives and their variants to the same in the target domain. This additional dictionary, hence, mixes the original training task with word translation (Mikolov et al., 2013b). Without any specific context, we hope the model can utilize the general cross-lingual supervised signals as anchor points so as to launch the semantic linking. We want to point out that deductive learning is partially different from *deductive reasoning*. Although there is an overlap, it is not necessary for the former to extract rules from observations like the inference conducted by the latter. In this work, we care more about the learning outcomes, rather than the reasoning process, through empirical evaluations. In practice, given the same example above, we directly make use of primitive "*jump*" and its variant "*jump_0*" as the source sequences, as well as the action "JUMP" as their identical target sequences. Words and phrases can be treated as text sequences of relatively short length. By exposing both the primitive rule "*jump*" $\rightarrow$ "JUMP" and the variants rule "*jump_0*"$\rightarrow$ "JUMP" during training, we aim to build the semantic connections between "*jump*" and "*jump_0*" deductively.

3 Systematic Generalization

The following section specifies the setup and outcome of the experiments. We first employ SCAN as the initial testbed to reproduce the one-shot generalization conditioned on the semantic linking. Then, we examine neural networks' potential to achieve this on SCAN and two real-world tasks of SP, followed by a sensitivity analysis.

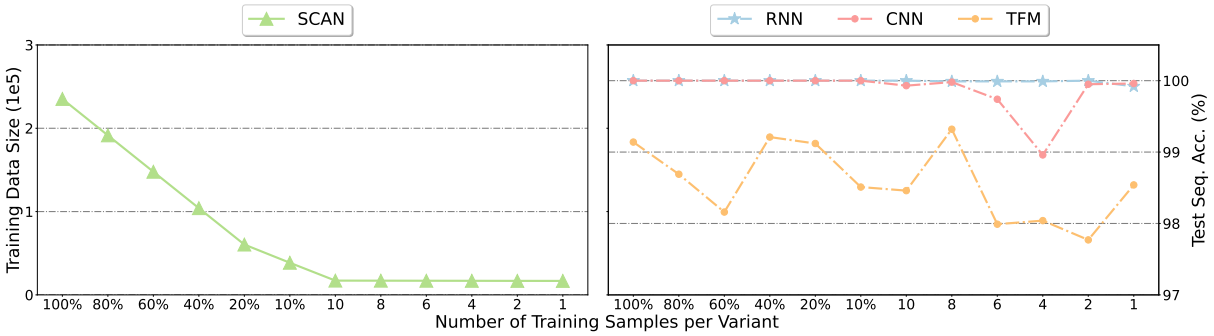

Figure 3: Experiments on SCAN expressing the total training size (left) and the test sequence accuracy (right) when the number of training samples per variant decreases from the complete set (100%) to a single sample (1).

3.1 Datasets

Some suggest SCAN is not enough to fully verify compositionality (Bastings et al., 2018; Keysers et al., 2020; Dankers et al., 2022a). Thus, we introduce GEO and ADV generated respectively from real SP datasets: Geography and Advising.[2] Example inputs and outputs can be found in Table 6.

SCAN (Lake and Baroni, 2018) is a diagnostic dataset proposed to investigate neural networks' compositionality.[3] It includes 20,910 pairs of commands to their instructed actions such as the example in Figure 1. We select {"*jump*", "*look*", "*run*", "*walk*" } as 4 primitives to be in line with previous works. We focus on lexical variants and create them by adding a suffix that consists of an underline and a unique number. We control the size of the variants set by setting the upper limit of this number. An example variant of "*jump*" is "*jump_0*" and both mean the same action "JUMP".

Geography is a common SP dataset (Zelle and Mooney, 1996; Srinivasan et al., 2017), containing 880 examples of queries paired with corresponding expressions. It is later formatted to SQL language with variables in the target sequences (Finegan-Dollak et al., 2018). **GEO** is generated from Geography. We regard 4 of 9 annotated variables as hypernyms and keep them as they are in SQL sequences. The other variables are restored by entities from the source sequence accordingly. As a result, the overall data size is 618 after processing. We can make use of the "is-a" hypernymy relation for semantic linking. Specifically, we select {"*new york city*", "*mississippi rivier*", "*dc*", "*dover*" } as 4 primitives[4] with their variants consisting of entities as co-hyponyms sharing the same variable group with primitives. An example variant of "*new york city*" is "*houston city*" and both are in the same variable group "CITY_NAME".

Advising includes 4,570 questions on course information paired with SQL queries (Finegan-Dollak et al., 2018). **ADV** is generated from Advising. We treat 4 of 26 annotated variables as hypernyms. Precisely, we select { "*a history of american film*", "*aaron magid*", "*aaptis*", "*100*" } as 4 primitives with their variants as co-hyponyms sharing the same variables. For instance, "*advanced at ai techniques*" is a co-hyponym of "*a history of american film*" sharing the same variable "TOPIC".

3.2 Models and experimental setup

Models. After testing many adapted versions, we employ three dominant model candidates, that is, RNN, CNN, and TFM. In terms of RNN, we reproduce bi-directional recurrent networks (Schuster and Paliwal, 1997) with long short-term memory units (Hochreiter and Schmidhuber, 1997) and an attention mechanism (Bahdanau et al., 2015). We follow the convolutional seq2seq architecture presented by Gehring et al. (2017) with regard to CNN and the attention-based structure proposed by Vaswani et al. (2017) in the case of TFM. More details are in Appendix B.

Training. We apply the mini-batch strategy to sample 128 sequence pairs for each training step. We use Adam optimizer (Kingma and Ba, 2015) with an ℓ_2 gradient clipping of 5.0 (Pascanu et al., 2013) and a learning rate of $1e^{-4}$. We freeze the maximum training epoch at 320 for CNN and 640 for RNN and TFM. To prevent uncontrolled interference, we train all models from scratch instead of fine-tuning (Devlin et al., 2019). For the same reason, we break words by whitespace tokenization rather than subword modeling. So, we can guarantee that words are treated separately as distinct

[2] github.com/jkkummerfeld/text2sql-data

[3] github.com/brendenlake/SCAN

[4] We randomly select 4 primitives from GEO and ADV to align with SCAN.

Data	SCAN					GEO					ADV				
	Exp. IL			Exp. DL		Exp. IL			Exp. DL		Exp. IL			Exp. DL	
	Sta.	Dif.	Cha.	Sta.	Dif.	Sta.	Dif.	Cha.	Sta.	Dif.	Sta.	Dif.	Cha.	Sta.	Dif.
Train Size	20946	20942	20928	20950	20946	724	720	711	728	724	6038	6034	5969	6040	6036
Test Size	308240	308240	308240	308240	308240	21350	21350	21350	21350	21350	107614	107614	107614	107614	107614

Table 1: Dataset statistics for inductive learning (IL) and deductive learning (DL) across Standard (Sta.), Difficult (Dif.), and Challenging (Cha.) in Section 3.4.

tokens with completely different embeddings.

Evaluation. Token and sequence accuracy serve as two primary metrics. The former allows partial errors in a sequence, while the latter strictly does not. Every reported number, along with the standard deviation, is the mean of five runs.

3.3 Experiment: meaningful learning

Thanks to their incredible algebraic compositionality (Chomsky, 1956), humans can effectively capture the underlying semantic connections between new and old concepts and generalize the prior knowledge to novel combinations by meaningful learning (Ausubel, 1963). To investigate the extent to which models can do the same, we probe the models' compositionality by introducing semantic linking. It is reasonable to illustrate the function of semantic linking through an ablation study, while its missing will lead to an out-of-vocabulary (OOV) issue since there will be no sample to expose variants during training. Replacing variants with other tokens (e.g., "*[unk]*") goes against our intent to investigate the generalization from primitives to their variants. It also leads to an unfair comparison, where all the variants, for example, go to the same unknown token and cause poor test accuracy. Instead, we gradually remove training samples for each variant until the one-shot learning scenario. We hope to observe the presence of models' meaningful learning by measuring the corresponding performance loss.

Experimental setup. Following section 2.2, we make use of 40 variants for 4 primitives and produce a total of 329,190 samples, including both primitive and variant samples. We randomly split them into a training set (80%) and a test set (20%). The training set is further processed to remove samples having multiple variants to ensure that each variant occurs only once in each sample. Eventually, the training set contains 235,002 samples. Models directly trained on this full dataset serve as baselines. Then, to format a gradual transition from baselines to the meaningful learning, we train the same models on various datasets with a decreasing number of augmented samples for each variant until the one-shot learning setting. Besides, we use the variant rule "*jump_0*" $\rightarrow$ "JUMP" as the only training sample for "*jump_0*" in the end as a case of our deductive learning introduced in Section 2.3 and consider the rest as our inductive learning.

Results. As elaborated in Figure 3, the solid line (SCAN) in green denotes the total training data size against the decreasing number of training samples per variant. The dashed line in other colors denotes the test sequence accuracy against the same horizontal axis. RNN has no significant performance drop when the training size is reduced from 100% to 1. It still achieves 99.92% test sequence accuracy when there is only one training sample for each variant. The same happens for CNN and TFM. Despite a slight fluctuation, they keep the results almost consistent regardless of whether the number of training variant samples is all or 1. It is not necessary to augment the training set nearly 14 times from 16,736 to 235,002 to cover all the possible variant compositions. The participation of a single sample is able to launch semantic linking via either inductive learning (a variant sample) or deductive learning (a variant rule), thus enabling models to achieve one-shot generalization. We put two plots in one figure to emphasize such a surprising observation through the strong contrast.

3.4 Experiment: semantic linking injection

The following two experiments evaluate models' systematic generalization, particularly for prior knowledge and semantic linking. A sliding scale of difficulty is carefully designed by weakening these two factors according to the assumption that the greater the difficulty, the more compositional skills are required. We further validate our findings on GEO and ADV. We use the same evaluation protocol across different datasets in this section.

Taking the base dataset as prior knowledge, we replace the primitives in source sequences with their variants to generate novel compositions, as introduced in Section 2.2. So far, the produced variant samples are not in the training set but in the

Data	Model	Token Acc.%			Seq. Acc.%		
		Standard	Difficult	Challenging	Standard	Difficult	Challenging
SCAN	RNN	99.99 ± 0.03	99.89 ± 0.19	99.96 ± 0.02	99.95 ± 0.08	99.85 ± 0.08	99.80 ± 0.31
	CNN	99.96 ± 0.08	99.76 ± 0.54	98.89 ± 2.44	99.85 ± 0.34	99.52 ± 1.07	97.57 ± 5.24
	TFM	98.91 ± 0.78	98.90 ± 1.10	98.76 ± 0.85	97.35 ± 1.62	96.86 ± 2.64	96.38 ± 2.81
GEO	RNN	75.71 ± 8.42	75.69 ± 6.12	73.46 ± 3.05	44.95 ± 14.69	43.27 ± 13.47	36.77 ± 5.60
	CNN	87.99 ± 2.67	79.51 ± 6.03	77.40 ± 2.48	69.46 ± 5.78	51.20 ± 8.64	48.58 ± 3.40
	TFM	75.37 ± 7.84	75.11 ± 4.88	68.41 ± 4.76	45.93 ± 12.42	44.59 ± 9.76	36.93 ± 7.47
ADV	RNN	58.61 ± 6.18	59.74 ± 5.67	58.11 ± 5.82	36.18 ± 5.75	35.69 ± 6.05	35.45 ± 6.69
	CNN	57.83 ± 7.55	54.05 ± 5.74	53.66 ± 2.57	45.08 ± 9.32	42.14 ± 6.90	41.37 ± 4.04
	TFM	53.43 ± 2.80	51.51 ± 4.50	49.17 ± 2.58	42.59 ± 3.65	41.28 ± 4.35	38.88 ± 2.68

Table 2: Evaluation results over RNN, CNN, and TFM on SCAN, GEO, and ADV across Standard, Difficult, and Challenging in Section 3.4.1.

test set. Hence, variants exist as OOV now. Then, we either incorporate one variant sample to introduce variants in training inductively or one variant rule to do so deductively. In the one-shot learning scenario, we ensure each variant only has a single sample and appears only once during training. For convenience, we keep the same settings for each primitive to have 10 variants in SCAN and a full variant set in GEO (e.g., 39 variants for "*new york city*"). It is noted in ADV that we randomly sample 5 variants for each primitive so that we cover all the variants with an appropriate test size.

3.4.1 Inductive learning

Experimental setup. We increase the difficulty by excluding primitive samples from the training set. It is worth noting that models have to generalize to not only new concepts but also their new compositions with a higher level of difficulty.

- **Standard**: Models are trained on prior knowledge and one variant sample per variant.
- **Difficult**: We remove from the prior knowledge primitive samples sharing the same context with their variant samples. For example, we remove "*jump twice*" due to "*jump_0 twice*", and thus models have to generalize to "*jump_0 twice*" without seeing "*jump twice*".
- **Challenging:** We also exclude from the prior knowledge primitive samples of the same length as their variant samples. For instance, models have to reproduce the same generalization to "*jump_0 twice*" without seeing primitive samples of length 2, including "*jump twice*", "*jump right*", "*jump left*", to name a few.[5]

SCAN. What stands out in Table 2 is an excellent one-shot generalization for all three networks.

[5]We remove samples that will not lead to unknown tokens.

Data	Model	Token Acc.%		Seq. Acc.%	
		Standard	Difficult	Standard	Difficult
SCAN	RNN	99.48 ± 0.71	98.70 ± 0.92	98.27 ± 2.38	95.39 ± 2.72
	CNN	99.99 ± 0.01	98.59 ± 3.10	99.96 ± 0.03	96.66 ± 7.27
	TFM	96.90 ± 1.78	96.68 ± 2.21	91.94 ± 4.04	91.26 ± 5.80
GEO	RNN	54.44 ± 7.15	39.71 ± 18.38	13.61 ± 7.08	7.76 ± 5.34
	CNN	41.86 ± 3.38	41.07 ± 7.48	4.85 ± 4.66	4.04 ± 2.18
	TFM	67.02 ± 6.91	65.97 ± 5.17	36.38 ± 10.08	31.57 ± 7.42
ADV	RNN	36.50 ± 7.66	36.42 ± 7.39	12.84 ± 4.31	12.66 ± 5.19
	CNN	43.51 ± 11.31	35.34 ± 14.68	32.33 ± 12.93	23.58 ± 16.04
	TFM	56.82 ± 3.79	53.33 ± 3.85	47.43 ± 3.71	43.24 ± 5.14

Table 3: Evaluation results over RNN, CNN, and TFM on SCAN, GEO, and ADV across Standard and Difficult in Section 3.4.2.

The participation of variant samples induces a near-perfect generalization. Even the worst results obtained by TFM in Challenging are around 98.76% and 96.38% in terms of token and sequence accuracy. The outcomes confirm that networks can inductively learn semantic relations from the context after semantic linking. The disappearance of training samples in Difficult and Challenging causes a performance drop. This is well in line with the widely accepted belief in meaningful learning theory that prior knowledge matters to generalization.

GEO & ADV. The more apparent changes in metrics again verify that prior knowledge is essential. Either excluding primitive samples containing the same context or those of the same sequence length can produce a steep fall in the generalization. On GEO, CNN can lose an absolute sequence accuracy of 18.26% from Standard to Difficult, and that for TFM drops 7.66%. This upholds our argument that generalization via meaningful learning is inseparable from sufficient prior knowledge. The overall decline in performance can be attributed to the switch from toy sets to actual datasets since both GEO and ADV own a much more complex encoding and decoding space than SCAN. There-

fore, we conclude that both prior knowledge and semantic linking exert powerful effects upon the potential of models to generalize systematically.

3.4.2 Deductive learning

Experimental setup. We increase the difficulty of compositional learning by excluding primitive rules from the training set as follows:

- **Standard**: Models are trained on the prior knowledge, primitive rules, and variant rules.
- **Difficult**: We remove primitive rules from the training set. Consequently, semantic links are weakened and depend on variant rules only.

SCAN. By incorporating deductive semantic linking, all three networks attain satisfying compositional generalization as shown in Table 3. CNN achieves the highest 99.96% in Standard, while TFM takes the lowest 91.26% in Difficult with regard to sequence accuracy. We can see a consistent decline in accuracy when we undermine the semantic linking by removing primitive rules from the training set. The most significant sequence accuracy drop of 3.3% comes from CNN when the difficulty upgrades. However, in Difficult, even the lowest one is impressive as there is only one variant rule to introduce each variant during training.

GEO & ADV. There is a persistent performance loss because of the absence of primitive rules from the training set across models. Concretely in GEO, the grade of CNN declines from 32.33% in Standard to 23.58% in Difficult in terms of sequence accuracy. The causal role of semantic linking is also demonstrated by varying the difficulty. The difference between Standard and Difficult indicated that either concept rules and just variant rules can connect primitives with their variants semantically, though the former is better than the latter. Moreover, models appear to realize systematic generalization better in an inductive way. By comparing Table 2 with Table 3, we find that current black-box neural nets are more capable of exploring patterns from specific samples with context information rather than understanding knowledge from general rules in our experiments. This sheds light on why current machine learning is still highly data-driven and can hardly break through the bottleneck to conduct advanced logic reasoning as human beings.

3.5 Sensitivity analysis

Regarding deductive learning, we conduct sensitivity analysis with a varying number of primitives (#primitives) from {1,2,3,4} and that of variants per primitive (#variants) from {1,5,10,15,20} over RNN on SCAN. The experimental setup is borrowed from Standard in Section 3.4.2.

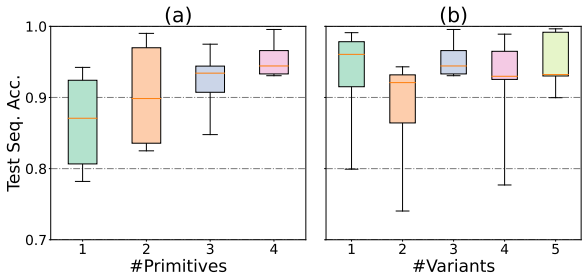

Figure 4: Experiments over RNN on SCAN with varying #primitives (a) and #variants (b).

Impact of #primitives. In Figure 4 (a), the generalization performance improves w.r.t. accuracy boosting and variance reduction when #primitives grows simultaneously. This is counter-intuitive as we thought primitive rules should work independently. A potential reason is semantic linking built by various *independent* primitive rules can profit each other to trigger a more robust and stable generalization. For example, "*jump*" $\rightarrow$ "JUMP" and "*look*" $\rightarrow$ "LOOK" may separate them from the context such as "***jump*** *right*" and "***look*** *right*". So, "*[concept] right*" functions as a compositional rule shared among primitive samples and finally encourages models to generalize more effectively.

Impact of #variants. As presented in Figure 4 (b), RNN generalizes consistently well when #variants goes up. Therefore, we report that the generalization among variants of the same primitive has a certain degree of independence within a reasonable range (e.g., #variants ≤ 20).

4 From SCAN to Real Data

Thus far, we have argued the feasibility of systematic generalization activated by semantic linking. We move on to discuss how it already benefits machines in solving real problems. Many recent papers propose to improve systematic generalization by techniques such as data augmentation (Andreas, 2020; Akyürek et al., 2021) and meta-learning (Lake, 2019; Conklin et al., 2021). The success is reasonable given our findings. Replacing fragments in real training samples with others that share similar contexts is supported by our inductive learning. We have demonstrated that similar context information can help establish the semantic links between new concepts and old ones, thus enabling models to generalize compositionally. By

Model	IWSLT'14				IWSLT'15			
	En-De		De-En		En-Fr		Fr-En	
	BLEU	SacreBLEU	BLEU	SacreBLEU	BLEU	SacreBLEU	BLEU	SacreBLEU
Baselines								
LSTM (Luong et al., 2015)	24.98	24.88	30.18	32.62	38.06	42.93	37.34	39.36
Transformer (Vaswani et al., 2017)	28.95	28.85	35.24	37.60	41.82	46.41	40.45	42.61
Dynamic Conv. (Wu et al., 2019)	27.39	27.28	33.33	35.54	40.41	45.32	39.61	41.42
+Vocabulary Augmentation								
LSTM (Luong et al., 2015)	$25.35\uparrow_{0.37}$	$25.38\uparrow_{0.50}$	$30.99\uparrow_{0.81}$	$33.63\uparrow_{1.01}$	$38.32\uparrow_{0.26}$	$43.30\uparrow_{0.37}$	$37.77\uparrow_{0.43}$	$39.83\uparrow_{0.47}$
Transformer (Vaswani et al., 2017)	$29.40\uparrow_{0.45}$	$29.29\uparrow_{0.44}$	$35.72\uparrow_{0.48}$	$38.07\uparrow_{0.47}$	$42.19\uparrow_{0.37}$	$46.68\uparrow_{0.27}$	$41.04\uparrow_{0.59}$	$43.15\uparrow_{0.54}$
Dynamic Conv. (Wu et al., 2019)	$27.60\uparrow_{0.21}$	$27.50\uparrow_{0.22}$	$33.62\uparrow_{0.29}$	$36.00\uparrow_{0.46}$	$40.87\uparrow_{0.46}$	$45.95\uparrow_{0.63}$	$39.95\uparrow_{0.34}$	$41.86\uparrow_{0.44}$

Table 4: Evaluation results over LSTM, Transformer, and Dynamic Conv. on IWSLT'14 En-De (English-German) and De-En, IWSLT'15 En-Fr (English-French) and Fr-En translations.

considering concepts as pointers in the memory, meta-learning equips models with memory loading to make connections between new and old concepts as semantic linking. The utility of similar unsupervised techniques (Xie et al., 2020) in both compositional generalization and real tasks can be attributed to inductive learning as well. Besides, our sensitivity analysis in Section 3.5 shows that adding seemingly independent primitive samples or rules can also improve the generalization, which has been further validated recently (Auersperger and Pecina, 2021; Patel et al., 2022).

In addition to inductive-based methods, some works (Mikolov et al., 2013b; Arthur et al., 2016; Nag et al., 2020), incorporating bilingual dictionaries in low-resource MT, can fall in the field of deductive-based ones. As a proof-of-concept, we reproduce the word-to-word augmentation, or called deductive learning in this work, by training models on not only the base training set but also concept rules. Intuitively, we wonder to which extent deductive semantic linking can promote models' performance in MT (IWSLT'14 and IWSLT'15) and SP (Geography and Advising). We report the evaluation results in Table 4 and Table 5. Details of models and data can be found in Appendix B and Appendix C.

4.1 Machine translation

Setup. We evaluate our approach on IWSLT'14 (Cettolo et al., 2014) English-German (En-De) and German-English (De-En), IWSLT'15 (Cettolo et al., 2015) English-French (En-Fr) and French-English (Fr-En) translation tasks. We follow the standard evaluation protocol (Ott et al., 2019) that keeps the original training set and validation set but combines multiple previous test sets for final evaluation. The test set of IWSLT'14 consists of IWSLT14.TED.dev{2010, 2012} and IWSLT14.TED.tst{2010, 2011, 2012}. That of IWSLT'15 includes IWSLT15.TED.tst{2014, 2015} (Ott et al., 2019). We apply BPE with 10K tokens for all tasks and report both BLEU (Papineni et al., 2002) and SacreBLEU (Post, 2018) scores for three baselines: LSTM (Luong et al., 2015), Transformer (Vaswani et al., 2017), and Dynamic Conv. (Wu et al., 2019) in comparsion with same structures augmented by our method.

Vocabulary augmentation. We introduce concept rules as *vocabulary augmentation* in MT. The semantic links between primitives and their variants can be built upon the synonymous relations between tokens such as "*heavily populated*" and "*populous*". From this, the source words paired with translated ones can be regarded as concept rules. It is noted that such relationships are reversible as shown in Figure 2, so a primitive can be a variant of the other primitive as well. In practice, we collect a dictionary of tokens in the source language and feed them to the Google Translation[6] so as to obtain a token map from the source language to the target one. The same operation can be repeated from the target language to the source one. Two dictionaries are combined into one with duplicates removed. Consequently, we get 144,874 token-level samples as a training supplementary for IWSLT'14 En-De and De-En, and 110,099 for IWSLT'15 En-Fr and Fr-En, which leads to a total of 305,113 training samples for IWSLT'14 En-De and De-En and 315,671 for IWSLT'15 En-Fr and Fr-En after such vocabulary augmentation.

Results. From Table 4, we observe a consistent improvement in both BLEU and SacreBLEU over all baselines after vocabulary augmentation, particularly up to 1 in SacreBlEU. The additional synonym pairs not only construct the semantic linking between tokens in two languages explicitly, but also create a complicated semantic linking network im-

[6] cloud.google.com/translate

Model	Geography				Advising			
	Train		Test		Train		Test	
	Token Acc.%	Seq. Acc.%	Token Acc.%	Seq. Acc.%	Token Acc.%	Seq. Acc.%	Token Acc.%	Seq. Acc.%
Baselines								
RNN	89.05	17.39	69.81	9.68	92.22	3.64	60.41	6.11
CNN	98.45	70.74	78.44	55.91	99.74	81.62	81.74	51.13
TFM	99.45	84.95	80.24	49.82	99.68	76.90	78.51	29.67
+Entity Augmentation								
RNN	87.47	29.96	$72.39\uparrow_{2.58}$	$15.05\uparrow_{5.37}$	88.82	30.97	$71.17\uparrow_{10.76}$	$16.06\uparrow_{9.95}$
CNN	97.54	76.03	$80.32\uparrow_{1.88}$	$60.93\uparrow_{5.02}$	99.65	87.01	$84.50\uparrow_{2.76}$	$56.02\uparrow_{4.89}$
TFM	99.30	85.73	$81.09\uparrow_{0.85}$	$54.84\uparrow_{5.02}$	99.57	86.94	$84.26\uparrow_{5.75}$	$35.08\uparrow_{5.41}$

Table 5: Evaluation results over RNN, CNN, and TFM on Geography and Advising.

plicitly because of synonyms within the single language and the transitivity nature of synonym relation. Our experiments prove that semantic linking, which allows models to generalize systematically, can be beneficial for improving MT performance.

4.2 Semantic parsing

Setup. We evaluate our method on two SP benchmarks, Geography, and Advising. We train the same models (i.e., RNN, CNN, and TFM) as we analyzed before without further hyperparameter tuning. There are some changes for CNN, where the learning rate is $5e^{-4}$ in Geography, and the maximum sequence length for the decoder position embedding is 312 in Advising. We split 10% training samples as the validation set to find the converged epoch and then add it back to the training set for the final report.

Entity augmentation. We introduce concept rules as *entity augmentation* in SP. The semantic links are established among co-hyponyms. We consider a variable as a hypernym for its values. By that, entities belonging to the same variable are co-hyponyms. Thus, we can regard entity values as primitives and the translations from primitives (e.g., "*new york city*") to their variables (e.g., "CITY_NAME") as primitive rules. To be specific, We construct entity dictionaries by collecting entities such as "*new york city*". They are translated to themselves since they do not change from the source natural language to the target SQL. For a fair comparison, a token from this extra dataset will be marked as a unique unknown mark, "*[unk]*", if it does not exist in the original base training set. After that, we have a map of 103 entity translations for Geography and 1846 for Advising, resulting in a training size change from 701 to 804 for Geography and from 3814 to 5660 for Advising.

Results. As elaborated in Table 5, all three networks can achieve better performance in terms of both accuracy and variance. A 10.76% token accuracy and 9.95% sequence accuracy boosting are observed from RNN on Advising after such entity augmentation. The results suggest that models can learn semantic linking or be more familiar with similar contexts from those primitive rules in a deductive way to enhance model systematic generalization and finally lead to better outcomes.

5 Conclusion

We revisit systematic generalization from a meaningful learning perspective. According to the theory, we conduct semantic linking to expose semantic relations between new and old concepts via either inductive learning or deductive learning. Experimental results on SCAN, GEO, and ADV support that seq2seq neural networks, as a class of modern machine learning methods, can behave systematically after semantic linking. Testing with various difficulties indicates that both semantic linking and prior knowledge are two essential factors in such generalization, in agreement with what humans do in meaningful learning. Finally, we group recent methods in either the inductive-based or deductive-based category, followed by a proof-of-concept, to highlight the already-existing advantages of meaningful learning in applications such as machine translation and semantic parsing.

We want to underline that, to the best of our knowledge, this work is the first one exploring the optimistic results observed by Lake and Baroni (2018). Our positive findings oppose the recent prevailing view that neural networks appear inherently ineffective in such cognitive capacity, thus confirming the mixed picture. By rationalizing recent findings from a meaningful learning perspective, we hope to encourage followers to interpret the exceptional generalization ability through the connection between neural nets and human cognition.

Limitations

We establish semantic relationships between primitives and their variants by either inductive or deductive learning. The incorporation of both learning skills is worth exploring further. We primarily utilize data augmentation techniques to expose the semantic information to models. Apart from that, there should be many other methods to achieve the same goal. Which method is most appropriate to realize semantic linking remains an open topic. Meanwhile, the application of meaningful learning to promote systematic generalization in practice (e.g., MT and SP) could have been expanded.

Acknowledgements

We gratefully appreciate Yewen Pu, Guan (Royal) Wang, Yichen Gong, Rong Zhang, and Hui Xue for sharing their pearls of wisdom. We also would like to express our special thanks of graitude to Yingying Huo for the support, as well as BlackboxNLP anonymous reviewers for their constructive feedback. This work was supported by Shining Lab, Learnable, Inc., and Alibaba Group.

References

Ekin Akyürek, Afra Feyza Akyürek, and Jacob Andreas. 2021. Learning to recombine and resample data for compositional generalization. In *International Conference on Learning Representations*.

Ekin Akyurek and Jacob Andreas. 2021. Lexicon learning for few shot sequence modeling. 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)*, pages 4934–4946, Online. Association for Computational Linguistics.

Jacob Andreas. 2020. Good-enough compositional data augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7556–7566, Online. Association for Computational Linguistics.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas. Association for Computational Linguistics.

Michal Auersperger and Pavel Pecina. 2021. Solving SCAN tasks with data augmentation and input embeddings. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 86–91, Held Online. INCOMA Ltd.

D.P. Ausubel. 1963. *The Psychology of Meaningful Verbal Learning*. Grune & Stratton.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Timothy Baldwin, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135, Beijing, China. Association for Computational Linguistics.

Fabian Barteld. 2017. Detecting spelling variants in non-standard texts. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 11–22, Valencia, Spain. Association for Computational Linguistics.

Joost Bastings, Marco Baroni, Jason Weston, Kyunghyun Cho, and Douwe Kiela. 2018. Jump to better conclusions: SCAN both left and right. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 47–55, Brussels, Belgium. Association for Computational Linguistics.

Philémon Brakel and Stefan Frank. 2009. Strong systematicity in sentence processing by simple recurrent networks. In *31th Annual Conference of the Cognitive Science Society (COGSCI-2009)*, pages 1599–1604. Cognitive Science Society.

M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, R. Cattoni, and Marcello Federico. 2015. The iwslt 2015 evaluation campaign. In *IWSLT*.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th iwslt evaluation campaign, iwslt 2014. In *Proceedings of the International Workshop on Spoken Language Translation, Hanoi, Vietnam*, volume 57.

Rahma Chaabouni, Roberto Dessì, and Eugene Kharitonov. 2021. Can transformers jump around right in natural language? assessing performance transfer from SCAN. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 136–148, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinyun Chen, Chen Liang, Adams Wei Yu, Dawn Song, and Denny Zhou. 2020. Compositional generalization via neural-symbolic stack machines. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Noam Chomsky. 1956. Syntactic structures.

Henry Conklin, Bailin Wang, Kenny Smith, and Ivan Titov. 2021. Meta-learning to compositionally generalize. 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)*, pages 3322–3335, Online. Association for Computational Linguistics.

Róbert Csordás, Kazuki Irie, and Jürgen Schmidhuber. 2021. The devil is in the detail: Simple tricks improve systematic generalization of transformers. In *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Punta Cana, Dominican Republic.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2022a. The paradox of the compositionality of natural language: A neural machine translation case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175, Dublin, Ireland. Association for Computational Linguistics.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022b. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Sarthak Dash, Md Faisal Mahbub Chowdhury, Alfio Gliozzo, Nandana Mihindukulasooriya, and Nicolas Rodolfo Fauceglia. 2020. Hypernym detection using strict partial order networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7626–7633.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Catherine Finegan-Dollak, Jonathan K. Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. 2018. Improving text-to-SQL evaluation methodology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 351–360, Melbourne, Australia. Association for Computational Linguistics.

Jerry A Fodor and Ernest Lepore. 2002. *The compositionality papers*. Oxford University Press.

Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Stefan L Frank, Willem FG Haselager, and Iris van Rooij. 2009. Connectionist semantic systematicity. *Cognition*, 110(3):358–379.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR.

Jonathan Gordon, David Lopez-Paz, Marco Baroni, and Diane Bouchacourt. 2020. Permutation equivariant models for compositional generalization in language. In *International Conference on Learning Representations*.

Robert F Hadley. 1994. Systematicity in connectionist language learning. *Mind & Language*, 9(3):247–272.

Hector Hammerly. 1975. The deduction/induction controversy. *The Modern Language Journal*, 59(1/2):15–18.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: how do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21.

Yichen Jiang and Mohit Bansal. 2021. Inducing transformer's compositional generalization ability via auxiliary sequence prediction tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6253–6265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2020. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Yoon Kim. 2021. Sequence-to-sequence learning with latent neural grammars. *Advances in Neural Information Processing Systems*, 34.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

William Labov. 1963. The social motivation of a sound change. *Word*, 19(3):273–309.

Brenden M Lake. 2019. Compositional generalization through meta sequence-to-sequence learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2879–2888. PMLR.

Yafu Li, Yongjing Yin, Yulong Chen, and Yue Zhang. 2021. On compositional generalization of neural machine translation. 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)*, pages 4767–4780, Online. Association for Computational Linguistics.

Yuanpeng Li, Liang Zhao, Jianyu Wang, and Joel Hestness. 2019. Compositional generalization for primitive substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4293–4302, Hong Kong, China. Association for Computational Linguistics.

Qian Liu, Shengnan An, Jian-Guang Lou, Bei Chen, Zeqi Lin, Yan Gao, Bin Zhou, Nanning Zheng, and Dongmei Zhang. 2020. Compositional generalization by learning analytical expressions. In *Advances in Neural Information Processing Systems*, volume 33, pages 11416–11427. Curran Associates, Inc.

Joao Loula, Marco Baroni, and Brenden Lake. 2018. Rearranging the familiar: Testing compositional generalization in recurrent networks. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 108–114, Brussels, Belgium. Association for Computational Linguistics.

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

John Lyons and Lyons John. 1995. *Linguistic semantics: An introduction*. Cambridge University Press.

Gary F Marcus. 1998. Rethinking eliminative connectionism. *Cognitive psychology*, 37(3):243–282.

Gary F Marcus. 2018. *The algebraic mind: Integrating connectionism and cognitive science*. MIT press.

Richard E Mayer. 2002. Rote versus meaningful learning. *Theory into practice*, 41(4):226–232.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Richard Montague et al. 1970. Universal grammar. *1974*, pages 222–46.

Sreyashi Nag, Mihir Kale, Varun Lakshminarasimhan, and Swapnil Singhavi. 2020. Incorporating bilingual dictionaries for low resource semi-supervised neural machine translation. In *International Conference on Learning Representations, Learning with Limited Labeled Data*.

Dong Nguyen and Jack Grieve. 2020. Do word embeddings capture spelling variation? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 870–881, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maxwell Nye, Armando Solar-Lezama, Josh Tenenbaum, and Brenden M Lake. 2020. Learning compositional rules via neural program synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 10832–10842. Curran Associates, Inc.

Peter Akinsola Okebukola and Olugbemiro J Jegede. 1988. Cognitive preference and learning mode as determinants of meaningful learning through concept mapping. *Science Education*, 72(4):489–500.

Inbar Oren, Jonathan Herzig, Nitish Gupta, Matt Gardner, and Jonathan Berant. 2020. Improving compositional generalization in semantic parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2482–2495, Online. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–1310–III–1318. JMLR.org.

Arkil Patel, Satwik Bhattamishra, Phil Blunsom, and Navin Goyal. 2022. Revisiting the compositional generalization abilities of neural sequence models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 424–434, Dublin, Ireland. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Lutz Prechelt. 1998. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Constance Shaffer. 1989. A comparison of inductive and deductive approaches to teaching foreign languages. *The Modern Language Journal*, 73(4):395–403.

Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional generalization and natural language variation: Can a semantic parsing approach handle both? 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)*, pages 922–938, Online. Association for Computational Linguistics.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4506–4515, Hong Kong, China. Association for Computational Linguistics.

Iyer Srinivasan, Konstas Ioannis, Cheung Alvin, Krishnamurthy Jayant, and Zettlemoyer Luke. 2017. Learning a neural semantic parser from user feedback. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 963–973.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Maja Stanojević et al. 2009. Cognitive synonymy: A general overview. *FACTA UNIVERSITATIS-Linguistics and Literature*, 7(2):193–200.

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

Scott Thornbury. 1999. *How to teach grammar*, volume 3. Longman Harlow.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Chengyu Wang and Xiaofeng He. 2020. Birre: learning bidirectional residual relation embeddings for supervised hypernymy detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3630–3640.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*.

Pengcheng Yin, Hao Fang, Graham Neubig, Adam Pauls, Emmanouil Antonios Platanios, Yu Su, Sam Thomson, and Jacob Andreas. 2021. Compositional generalization for neural semantic parsing via span-level supervised attention. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2810–2823, Online. Association for Computational Linguistics.

Jiale Yu, Yongliang Shen, Xinyin Ma, Chenghao Jia, Chen Chen, and Weiming Lu. 2020a. Synet: Synonym expansion using transitivity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1961–1970.

Yue Yu, Yinghao Li, Jiaming Shen, Hao Feng, Jimeng Sun, and Chao Zhang. 2020b. Steam: Self-supervised taxonomy expansion with mini-paths. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1026–1035.

John M. Zelle and Raymond J. Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, pages 1050–1055.

A Semantic Links

Lexical Variant refers to an alternative expression form for the same concept, where the various forms may derive from foreign languages, abbreviations, and even mistakes. A basic assumption is that all languages change over time due to non-linguistic factors. Since the rise of sociolinguistics in the 1960s, studies on linguistic variability, a characteristic of language, are central to the language use and motivations for speakers to vary the pronunciation, word choice, or morphology of existing concepts (Labov, 1963). Taking "*United States of America*" as an example, people have generally accepted the semantic connections among its lexical variants in history, including "*America*" and "*United States*", as well as the initialisms "*U.S.*" and "*U.S.A*". Many efforts have been devoted on lexical variants representation (Nguyen and Grieve, 2020), detection (Barteld, 2017), normalization (Baldwin et al., 2015) to keep machines up with the trend of the times.

Co-hyponym is a linguistic term to designate a semantic relation between two group members belonging to the same broader class, where each member is a hyponym, also called subtype or subordinate, and the class is a hypernym (Lyons and John, 1995). The "is-a" hypernymy relation between a generic hypernym and its specific hyponyms builds semantic connections among co-hyponyms. An example of such a hierarchical structure can be "*Mississippi*" and "*California*" in the domain of "*state*". Specifically, "*Mississippi*" and "*California*" are two hyponyms, and "*state*" is a hypernym. Thus, "*Mississippi*" and "*California*" are semantically connected to be co-hyponyms for each other. Harvesting hypernymy relations (Wang and He, 2020) plays an essential role for downstream knowledge graph construction (Ji et al., 2021), out-vocabulary generalization (Dash et al., 2020), and taxonomy expansion (Yu et al., 2020b).

Synonym stands for a word, morpheme, or phrase that shares exactly or nearly the same semantics with another one. Many tend to assume synonyms are utterances that occur in most contexts in common, so they are semantically closely related enough to be synonyms for each other (Rubenstein and Goodenough, 1965; Harris, 1954). The existence of the association to contexts is a basic assumption supporting the advance of recent masked language modeling (Devlin et al., 2019). Given that, one of the definitions of a synonymous relation is a semantic link between two expressions if substitution of one for the other never hurts the true value of the context (Stanojević et al., 2009). For instance, the substitution of "*heavily populated*" for "*populous*" will seldom alter the truth of the sentence in Figure 2. Such semantic similarity can be observed in continuous vector space from a trained representation as well (Mikolov et al., 2013a). Synonym discovery (Yu et al., 2020a) has been a fundamental job to construct knowledge base and thus benefits substantial researches.

B Models

All models are built within the encoder-decoder framework (Sutskever et al., 2014). We reproduce RNN, CNN, and TFM by ourselves to have fewer parameters than the original versions for the experimental purposes. The dropout rate is 0.5 for RNN, CNN, and TFM (Srivastava et al., 2014). We implement LSTM, Transformer, and Dynamic Conv. within the library *fairseq*.[7] (Ott et al., 2019) and inherit its default model structures.[8] In contrast to early stopping (Prechelt, 1998), we prefer a fixed training regime sufficient enough for models to fully converge in practice with a focus on the systematic generalization observation instead of superior structure exploration. Training is on a single Nvidia Tesla V100. Without specific notes, hyperparameters are shared throughout the work.

RNN denotes bi-directional recurrent network (Schuster and Paliwal, 1997; Hochreiter and Schmidhuber, 1997) with long short-term memory

[7] https://github.com/pytorch/fairseq

[8] LSTM is adapted from *lstm_luong_wmt_en_de*; Transformer is adapted from *transformer_iwslt_de_en*; Dynamic Conv. is adapted from *lightconv_iwslt_de_en*.

Data	Sequence	
SCAN	Source	*jump twice*
	Target	JUMP JUMP
GEO	Source	*how many people in new york city*
	Target	SELECT CITY alias0 . POPULATION FROM CITY AS CITY alias0 WHERE CITY alias0 . CITY_NAME = CITY_NAME ;
ADV	Source	*Which department includes a history of american film ?*
	Target	SELECT DISTINCT COURSE alias0 . DEPARTMENT FROM COURSE AS COURSE alias0 WHERE COURSE alias0 . NAME LIKE TOPIC ;
Geography	Source	*how many people live in new york*
	Target	SELECT STATE alias0 . POPULATION FROM STATE AS STATE alias0 WHERE STATE alias0 . STATE_NAME = " new york " ;
Advising	Source	*I would like to see A History of American Film courses of 2 credits .*
	Target	SELECT DISTINCT COURSE alias0 . DEPARTMENT , COURSE alias0 . NAME , COURSE alias0 . NUMBER FROM COURSE AS COURSE alias0 WHERE (COURSE alias0 . DESCRIPTION LIKE "% A History of American Film %" OR COURSE alias0 . NAME LIKE "% A History of American Film %") AND COURSE alias0 . CREDITS = 2 ;

Table 6: Example source and target sequences from SCAN, GEO, ADV, Geography, and Advising.

Data	Primitive	Semantic Links	Variant	Concept Rule	
				Primitive Rule	Variant Rule
SCAN	*jump*	Lexical Variant	*jump_0*	*jump* → JUMP	*jump_0* → JUMP
	look		*look_0*	*look* → LOOK	*look_0* → LOOK
	run		*run_0*	*run* → RUN	*run_0* → RUN
	walk		*walk_0*	*walk* → WALK	*walk_0* → WALK
GEO	*new york city*	Co-hyponym	*houston city*	*new york city* → CITY_NAME	*houston city* → CITY_NAME
	mississippi rivier		*red rivier*	*mississippi rivier* → RIVER_NAME	*red rivier* → RIVER_NAME
	dc		*kansas*	*dc* → STATE_NAME	*kansas* → STATE_NAME
	dover		*salem*	*dover* → CAPITAL_NAME	*salem* → CAPITAL_NAME
ADV	*a history of american film*	Co-hyponym	*advanced ai techniques*	*a history of american film* → TOPIC	*advanced ai techniques* → TOPIC
	aaron magid		*cargo*	*aaron magid* → INSTRUCTOR	*cargo* → INSTRUCTOR
	aaptis		*survmeth*	*aaptis* → DEPARTMENT	*survmeth* → DEPARTMENT
	100		*171*	*100* → NUMBER	*171* → NUMBER

Table 7: Concept rules with primitives and their example variants.

units and an attention mechanism (Bahdanau et al., 2015). Its encoder consists of two layers with a hidden size of 256 in each direction, and its decoder has one layer with a hidden size of 512. The embedding size is 512 for both encoder and decoder. There are a total of $5.29M$ trainable parameters. Teacher forcing with a rate of 0.5 serves to spur up the training process (Williams and Zipser, 1989).

CNN denotes the fully convolutional seq2seq network (Gehring et al., 2017). The size of the position embedding layer is 128 for encoding and 256 for decoding, while that of the token embedding layer is 512 for both encoding and decoding. There are 10 convolutional layers with 512 as the hidden size and 3 as the kernel size in both encoder and decoder, resulting in a total of $33.55M$ trainable parameters.

TFM denotes transformers, an attention-based network (Vaswani et al., 2017). As a tiny version, TFM has 2 layers for each encoder and decoder with 8 attention heads and a dimension of 512. The size of the feedforward layer is 2048. We utilize the cyclic nature of $\sin$ and $\cos$ functions to represent token positions. There are a total of $15.02M$ trainable parameters.

LSTM is adapted from the recurrent network used by Luong et al. (2015) for statistical MT. The size of the embedding layer is 1000. There are 4 layers in both encoder and decoder with a hidden size of 512 and a dropout rate of 0.2.

Transformer, the same as TFM, is adapted from the base version of transformers in the work of Vaswani et al. (2017), while TFM is a tiny version to test systematic generalization. The dimension is 512 for the embedding layer, 1024 for the feedforward layer, and 512 for the attention layer. There are 6 attention blocks in both encoder and decoder with 4 attention heads and 0.3 dropout probability.

Dynamic Conv. is adapted from the seq2seq convolutional network proposed by Wu et al. (2019), where the hidden size of the embedding layer, encoder layer, and decoder layer is 512. The number of attention heads is 4, and the dimension of the feedforward layer is 1024 for both encoder and encoder. There are 6 layers in the encoder and 7 layers in the decoder. The dropout rate is 0.1 for both attention and weight units.

C Data

IWSLT involves IWSLT'14 (Cettolo et al., 2014) English-German (En-De) and German-English (De-En), IWSLT'15 (Cettolo et al., 2015) English-French (En-Fr) and French-English (Fr-En) translation tasks. The goal is to translate a sen-

Data	Primitive	Variant	#Variants	Prompt
SCAN	*jump*	*jump_0*	10	*[concept] twice*
GEO	*new york city*	*houston city*	39	*how many people in [concept]*
	mississippi rivier	*red rivier*	9	*how long is [concept]*
	dc	*kansas*	49	*where is [concept]*
	dover	*salem*	8	*what states capital is [concept]*
ADV	*a history of american film*	*advanced ai techniques*	5/424	*who teaches [concept] ?*
	aaron magid	*cargo*	5/492	*does [concept] give upper-level courses ?*
	aaptis	*survmeth*	5/1720	*name core courses for [concept] .*
	100	*171*	5/1895	*can undergrads take [concept] ?*

Table 8: Prompts with example primitives and sampled variants. In SCAN, primitives share the same prompt and the number of variants can be changed. In GEO, we make use of the full variants set. In ADV, we randomly sample 5 variants for each source sequence so that we cover all the variants with a test set of an appropriate size. We generate variant samples by filling the prompt with variants accordingly.

Data		SCAN					GEO					ADV					Geography		Advising	
		Exp. 1			Exp. 2		Exp. 1			Exp. 2		Exp. 1			Exp. 2		Bas.	Aug.	Bas.	Aug.
		Sta.	Dif.	Cha.	Sta.	Dif.	Sta.	Dif.	Cha.	Sta.	Dif.	Sta.	Dif.	Cha.	Sta.	Dif.				
Train Size		20946	20942	20928	20950	20946	724	720	711	728	724	6038	6034	5969	6040	6036	598	701	3814	5660
Test Size		308240	308240	308240	308240	308240	21350	21350	21350	21350	21350	107614	107614	107614	107614	107614	279	279	573	573
Time	RNN	21					5					19					4	5	27	35
	CNN	17					1.2					11					1	1.2	12	19
	TFM	7					0.5					5					0.4	0.5	6	8

Table 9: Data statistics and training time per epoch in seconds. The batch size of each epoch for GEO and Geography is 32, and that for the others is 128.

tence from one language to the other. The IWSLT'14 En-De and De-EN have 160,239 sequence pairs for training and 7,283 for validation. We make use of IWSLT14.TED.dev{2010, 2012} and IWSLT14.TED.tst{2010, 2011, 2012} to measure translation performance, resulting in a total of 6,750 test samples. In terms of IWSLT'15 En-Fr and Fr-En, there are 205,572 sequence pairs for training. We employ IWSLT15.TED.dev2010 and IWSLT15.TED.tst{2010, 2011, 2012, 2013} as the validation set and IWSLT15.tst{2014, 2015} as the test set. As a consequence, there are 5,519 samples for validation and 2,385 for evaluation. For all four translation tasks, we apply BPE with 10K tokens to share.

D Experiments

D.1 Inductive learning

Semantic linking can be operated via inductive learning, where we replace the concept in the prompt with primitives and their variants. The learning rate to train CNN in GEO is changed to $5e^{-4}$. Prompts used in SCAN, GEO, and ADV are expressed in Table 8. Detailed experimental results with respect to three levels can be found in Table 10, Table 11, and Table 12.

D.2 Deductive learning

Semantic linking can be established via deductive learning, where we put concept rules without context information in the training set instead of specific sequence samples. Example concept rules for SCAN, GEO, and ADV are presented in Table 7. Detailed experimental results with respect to two levels can be found in Table 13 and Table 14.

D.3 Sensitivity analysis

In sensitivity analysis, we adjust the number of primitives (#primitives) and the number of variants per primitive (#variants) over RNN on SCAN. The complete versions of Figure 4 in Section 3.5 are presented as Figure 5 and Figure 6 for #primitives and #variants respectively.

Data	Model	Train			Test		
		Loss	Token Acc.%	Seq. Acc.%	Loss	Token Acc.%	Seq. Acc.%
SCAN	RNN	0.00 ± 0.00	100.00 ± 0.00	99.99 ± 0.02	0.00 ± 0.00	99.99 ± 0.03	99.95 ± 0.08
	CNN	0.00 ± 0.00	99.81 ± 0.09	98.78 ± 0.55	0.00 ± 0.00	99.96 ± 0.08	99.85 ± 0.34
	TFM	0.00 ± 0.00	99.82 ± 0.02	98.83 ± 0.12	0.06 ± 0.03	98.91 ± 0.78	97.35 ± 1.62
GEO	RNN	0.15 ± 0.02	97.73 ± 0.42	80.25 ± 2.81	1.36 ± 0.48	75.71 ± 8.42	44.95 ± 14.69
	CNN	0.07 ± 0.01	98.23 ± 0.39	76.80 ± 2.25	9.01 ± 4.26	87.99 ± 2.67	69.46 ± 5.78
	TFM	0.02 ± 0.00	99.63 ± 0.07	91.60 ± 1.41	4.55 ± 1.39	75.37 ± 7.84	45.93 ± 12.42
ADV	RNN	0.03 ± 0.01	99.40 ± 0.13	82.74 ± 2.78	6.04 ± 0.95	58.61 ± 6.18	36.18 ± 5.75
	CNN	0.01 ± 0.01	99.59 ± 0.07	85.13 ± 1.95	23.56 ± 4.95	57.83 ± 7.55	45.08 ± 9.32
	TFM	0.00 ± 0.00	99.92 ± 0.01	96.14 ± 0.28	15.12 ± 1.00	53.43 ± 2.80	42.59 ± 3.65

Table 10: Results of Standard inductive learning.

Data	Model	Train			Test		
		Loss	Token Acc.%	Seq. Acc.%	Loss	Token Acc.%	Seq. Acc.%
SCAN	RNN	0.00 ± 0.00	100.00 ± 0.00	99.99 ± 0.01	0.00 ± 0.00	99.96 ± 0.02	99.85 ± 0.08
	CNN	0.00 ± 0.00	99.77 ± 0.19	98.62 ± 1.13	0.03 ± 0.06	99.76 ± 0.54	99.52 ± 1.07
	TFM	0.00 ± 0.00	99.79 ± 0.03	98.59 ± 0.12	0.06 ± 0.03	98.90 ± 1.10	96.86 ± 2.64
GEO	RNN	0.16 ± 0.03	97.39 ± 0.67	78.33 ± 4.31	1.29 ± 0.27	75.69 ± 6.12	43.27 ± 13.47
	CNN	0.07 ± 0.01	98.25 ± 0.13	76.53 ± 1.68	13.87 ± 3.19	79.51 ± 6.03	51.20 ± 8.64
	TFM	0.00 ± 0.11	99.60 ± 0.11	91.33 ± 1.46	4.50 ± 0.80	75.11 ± 4.88	44.59 ± 9.76
ADV	RNN	0.03 ± 0.01	99.26 ± 0.21	79.57 ± 4.12	5.80 ± 0.92	59.74 ± 5.67	35.69 ± 6.05
	CNN	0.02 ± 0.00	99.56 ± 0.05	84.06 ± 1.57	24.58 ± 3.40	54.05 ± 5.74	42.14 ± 6.90
	TFM	0.00 ± 0.00	99.91 ± 0.01	95.88 ± 0.23	15.84 ± 1.51	51.51 ± 4.50	41.28 ± 4.35

Table 11: Results of Difficult inductive learning.

Data	Model	Train			Test		
		Loss	Token Acc.%	Seq. Acc.%	Loss	Token Acc.%	Seq. Acc.%
SCAN	RNN	0.00 ± 0.00	100.00 ± 0.00	99.99 ± 0.02	0.20 ± 0.45	99.95 ± 0.08	99.80 ± 0.31
	CNN	0.00 ± 0.00	99.85 ± 0.05	99.00 ± 0.30	0.14 ± 0.31	98.89 ± 2.44	97.57 ± 5.24
	TFM	0.00 ± 0.00	99.82 ± 0.05	98.85 ± 0.27	0.07 ± 0.05	98.76 ± 0.85	96.38 ± 2.81
GEO	RNN	0.15 ± 0.04	97.76 ± 0.74	79.77 ± 4.19	1.52 ± 0.29	73.46 ± 3.05	36.77 ± 5.60
	CNN	0.07 ± 0.01	98.23 ± 0.17	75.98 ± 1.46	15.83 ± 4.56	77.40 ± 2.48	48.53 ± 3.40
	TFM	0.02 ± 0.00	99.60 ± 0.06	91.00 ± 1.20	6.01 ± 1.03	68.41 ± 4.76	36.93 ± 7.47
ADV	RNN	0.03 ± 0.01	99.23 ± 0.13	79.90 ± 1.85	5.95 ± 0.90	58.11 ± 5.82	35.45 ± 6.69
	CNN	0.01 ± 0.01	99.68 ± 0.15	87.90 ± 5.05	23.08 ± 6.34	53.66 ± 2.57	41.37 ± 4.04
	TFM	0.00 ± 0.00	99.93 ± 0.01	96.41 ± 0.24	16.59 ± 0.98	49.17 ± 2.58	38.88 ± 2.68

Table 12: Results of Challenging inductive learning.

Data	Model	Train			Test		
		Loss	Token Acc.%	Seq. Acc.%	Loss	Token Acc.%	Seq. Acc.%
SCAN	RNN	0.00 ± 0.00	99.99 ± 0.03	99.90 ± 0.23	0.05 ± 0.06	99.48 ± 0.71	98.27 ± 2.38
	CNN	0.00 ± 0.00	99.79 ± 0.14	98.78 ± 0.79	0.00 ± 0.00	99.99 ± 0.01	99.96 ± 0.03
	TFM	0.00 ± 0.00	99.82 ± 0.03	98.78 ± 0.17	0.27 ± 0.22	96.90 ± 1.78	91.94 ± 4.04
GEO	RNN	0.17 ± 0.03	97.50 ± 0.30	78.54 ± 2.16	2.83 ± 0.69	54.44 ± 7.15	13.61 ± 7.08
	CNN	0.08 ± 0.01	97.97 ± 0.24	77.03 ± 1.42	51.08 ± 25.97	41.86 ± 3.38	4.85 ± 4.66
	TFM	0.02 ± 0.00	99.54 ± 0.31	91.82 ± 2.27	6.03 ± 1.56	67.02 ± 6.91	36.38 ± 10.08
ADV	RNN	0.08 ± 0.02	98.64 ± 0.31	68.84 ± 4.57	7.95 ± 1.13	36.50 ± 7.66	12.84 ± 4.31
	CNN	0.02 ± 0.00	99.53 ± 0.07	84.64 ± 1.20	31.12 ± 4.76	43.51 ± 11.31	32.33 ± 12.93
	TFM	0.00 ± 0.00	99.91 ± 0.02	96.33 ± 0.37	13.72 ± 1.41	56.82 ± 3.79	47.43 ± 3.71

Table 13: Results of Standard deductive learning.

Data	Model	Train			Test		
		Loss	Token Acc.%	Seq. Acc.%	Loss	Token Acc.%	Seq. Acc.%
SCAN	RNN	0.00 ± 0.00	99.99 ± 0.01	99.95 ± 0.07	0.08 ± 0.08	98.70 ± 0.92	95.39 ± 2.72
	CNN	0.00 ± 0.00	99.62 ± 0.34	98.82 ± 1.09	0.13 ± 0.29	98.59 ± 3.10	96.66 ± 7.27
	TFM	0.00 ± 0.00	99.82 ± 0.03	98.78 ± 0.12	0.21 ± 0.20	96.68 ± 2.21	91.26 ± 5.80
GEO	RNN	0.20 ± 0.03	96.93 ± 0.71	75.35 ± 3.57	4.40 ± 2.50	39.71 ± 18.38	7.67 ± 5.34
	CNN	0.08 ± 0.01	97.77 ± 0.76	76.41 ± 2.80	32.94 ± 4.26	41.07 ± 7.48	4.04 ± 2.18
	TFM	0.02 ± 0.00	99.56 ± 0.11	91.08 ± 1.56	5.97 ± 1.05	65.97 ± 5.17	31.57 ± 7.42
ADV	RNN	0.08 ± 0.02	98.54 ± 0.28	67.10 ± 3.45	7.87 ± 1.01	36.42 ± 7.39	12.66 ± 5.19
	CNN	0.04 ± 0.05	98.78 ± 1.91	77.14 ± 23.28	32.44 ± 6.07	35.34 ± 14.68	23.58 ± 16.04
	TFM	0.00 ± 0.00	99.92 ± 0.02	96.41 ± 0.26	14.92 ± 1.31	53.33 ± 3.85	43.24 ± 5.14

Table 14: Results of Difficult deductive learning.

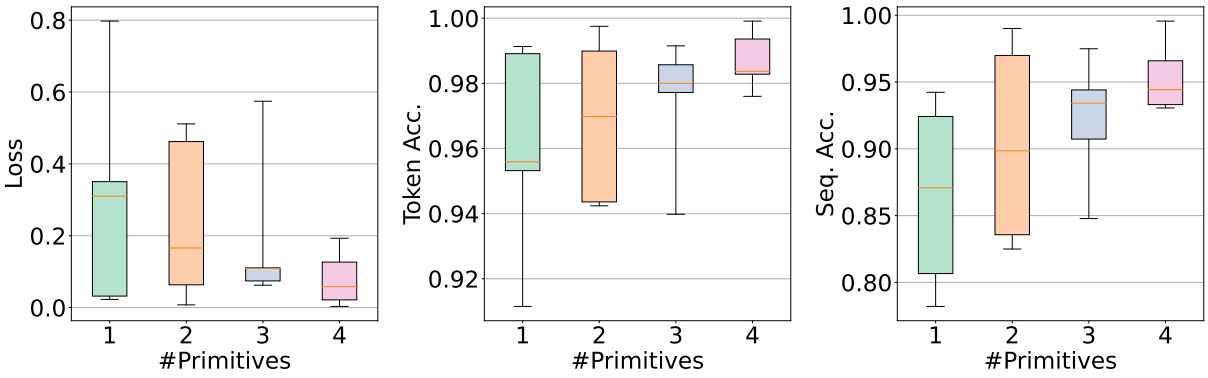

Figure 5: The complete version of Figure 4 in Section 3.5 regarding #primitives.

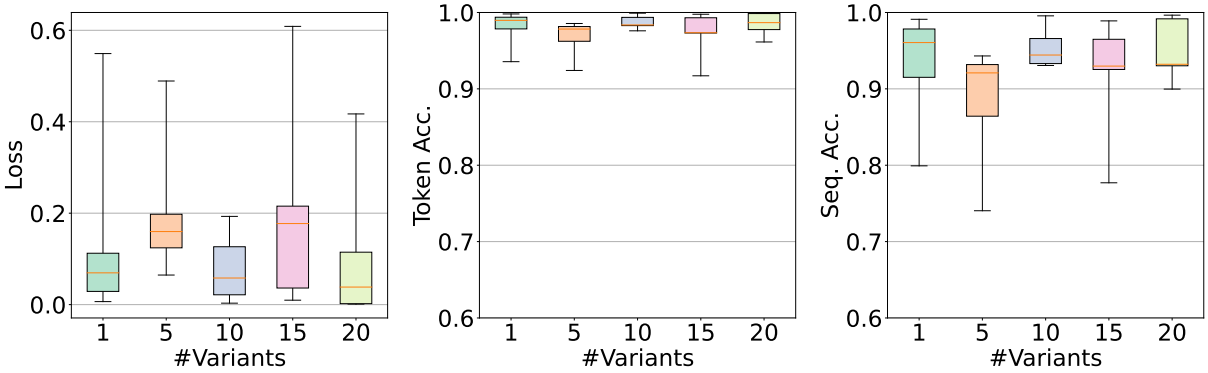

Figure 6: The complete version of Figure 4 in Section 3.5 regarding #variants.