On Compositional Generalization of Neural Machine Translation

Yafu Li[♠♡], **Yongjing Yin**[♠♡], **Yulong Chen**[♠♡], **Yue Zhang**[♡◇]
[♠] Zhejiang University
[♡] School of Engineering, Westlake University
[◇] Institute of Advanced Technology, Westlake Institute for Advanced Study
yafuly@gmail.com *yinyongjing@westlake.edu.cn*
yulongchen1010@gmail.com *yue.zhang@wias.org.cn*

Abstract

Modern neural machine translation (NMT) models have achieved competitive performance in standard benchmarks such as WMT. However, there still exist significant issues such as robustness, domain generalization, etc. In this paper, we study NMT models from the perspective of compositional generalization by building a benchmark dataset, CoGnition, consisting of 216k clean and consistent sentence pairs. We quantitatively analyze effects of various factors using compound translation error rate, then demonstrate that the NMT model fails badly on compositional generalization, although it performs remarkably well under traditional metrics.

1 Introduction

Neural machine translation (NMT) has shown competitive performance on benchmark datasets such as IWSLT and WMT (Vaswani et al., 2017; Edunov et al., 2018a; Liu et al., 2020a), and even achieves parity with professional human translation under certain evaluation settings (Hassan et al., 2018). However, the performance can be relatively low in out-of-domain and low-resource conditions. In addition, NMT systems show poor robustness and vulnerability to input perturbations (Belinkov and Bisk, 2018a; Cheng et al., 2019). One example is shown in Table 1, where simple substitution of a word yields translation with completely different semantics. Many of these issues origin from the fact that NMT models are trained end-to-end over large parallel data, where new test sentences can be sparse.

Disregarding out-of-vocabulary words, a main cause of sparsity is semantic composition: given a limited vocabulary, the number of possible compositions grows exponentially with respect to the composite length. The ability to understand and produce a potentially infinite number of novel combinations of known components, namely *compositional generalization* (Chomsky; Montague; Lake and Baroni, 2018; Keysers et al., 2020), has been demonstrated deficient in many machine learning (ML) methods (Johnson et al., 2017a; Lake and Baroni, 2018; Bastings et al., 2018; Loula et al., 2018; Russin et al., 2019a).

Input	Translation
Taylor **breaks** his promise	泰勒信守诺言 (Taylor keeps his promise)
James **breaks** his promise	詹姆斯违反诺言 (James breaks his promise)

Table 1: Translation samples obtained from one popular web translation engine on January 19, 2021.

In this paper, we study compositional generalization in the context of machine translation. For example, if "*red cars*" and "*blue balls*" are seen in training, a competent algorithm is expected to translate "*red balls*" correctly, even if the phrase has not been seen in training data. Intuitively, the challenge increases as the composite length grows. Recently, several studies have taken steps towards this specific problem. They either use a few dedicated samples (i.e., 8 test sentences) for evaluation (Lake and Baroni, 2018; Li et al., 2019b; Chen et al., 2020), or make simple modifications in sampled source sentences such as removing or adding adverbs, and concatenating two sentences (Raunak et al., 2019; Fadaee and Monz, 2020a). Such experimental data is limited in size, scope and specificity, and the forms of composition are coarse-grained and non-systematic. As a result, no qualitative conclusions have been drawn on the prevalence and characteristics of this problem in modern NMT.

We make a first large-scale general domain investigation, constructing the CoGnition dataset (**Co**mpositional **G**eneralizati**on** Machi**n**e **T**ranslat**ion** Dataset), a clean and consistent paral-

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 4767–4780
August 1–6, 2021.

Dataset	Type	Source	Target
SCAN	*Atoms* *Compounds*	jump, twice jump twice	JUMP JUMP
CFQ	*Atoms* *Compounds*	Who [predicate] [entity], directed, Elysium Who directed Elysium	SELECT DISTINCT ?x0 WHERE { ?x0 a ns:people.person . ?x0 ns:film.director.film m.0gwm_wy}
CoGnition	*Atoms* *Compounds* *Sentences*	the, doctor, he liked the doctor he liked The doctor he liked was sick	他喜欢的医生病了

Table 2: Examples of SCAN, CFQ, and our CoGnition datasets.

lel dataset in English-Chinese, along with a synthetic test set to quantify and analyze the compositional generalization of NMT models. In particular, we define frequent syntactic constituents as *compounds*, and basic semantic components in constituents as *atoms*. In addition to the standard training, validation and test sets, the CoGnition dataset contains a compositional generalization test set, which contains novel compounds in each sentence, so that both the generalization error rate can be evaluated, and its influence on BLEU (Papineni et al., 2002) can be quantified. Our compositional generalization test set consists of 2,160 novel compounds, with up to 5 atoms and 7 words. In this way, generalization ability can be evaluated based on compound translation error rate.

Empirical results show that the dominant Transformer (Vaswani et al., 2017) NMT model faces challenges in translating novel compounds, despite its competitive performance under traditional evaluation metrics such as BLEU. In addition, we observe that various factors exert salient effects on model's ability of compositional generalization, such as compound frequency, compound length, atom co-occurrence, linguistic patterns, and context complexity. The CoGnition dataset along with the automatic evaluation tool are realesed on https://github.com/yafuly/CoGnition.

2 Related Work

Analysis of NMT. Our work is related to research analyzing NMT from various perspectives. There has been much linguistic analysis of NMT representations (Shi et al., 2016; Belinkov et al., 2017; Bisazza and Tump, 2018), interpretability (Ding et al., 2017; He et al., 2019; Voita et al., 2019a), and attention weights (Voita et al., 2019b; Michel et al., 2019). Robustness is also an important research direction. Work has shown that NMT models are prone to be negatively affected by both synthetic and natural noise (Belinkov and Bisk, 2018b; Cheng et al., 2018; Ebrahimi et al., 2018). For better exploration of robust NMT, Michel and Neubig (2018) propose an MTNT dataset containing several types of noise. Wang et al. (2020) provide in-depth analyses of inference miscalibration of NMT resulting from the discrepancy between training and inference. Our work is in line but we discuss robustness from the perspective of compositional generalization.

In this respect, Lake and Baroni (2018) propose a simple experiment to analyze compositionality in MT, followed by Chen et al. (2020) and Li et al. (2019b). Specifically, they introduce a novel word "*dax*", and their training data contains a single pattern of sentence pairs (e.g. "*I am daxy*", "*je suis daxiste*") while the test set contains different patterns. However, their work is limited in that there are only 8 sentences in the test set. Raunak et al. (2019) observe a performance drop on a dataset of concatenated source sentences. Fadaee and Monz (2020b) modify source sentences by removing adverbs, substituting numbers, inserting words that tend to keep syntax correct (e.g. "*very*"), and changing the gender, and find unexpected changes in the translation. In contrast to these studies, we quantitatively measure compositionality of NMT under compound translation error rate.

Translation involves various challenges such as low-frequency words, polysemy and compositional complexity. In this work, we **focus** on how the NMT model generalizes to complex compositions in a controllable setting and minimize the effects of the other factors.

Compositional Generalization. Neural networks have been shown sample-inefficient, requiring large-scale training data, which suggests that they may lack compositionality (Lake and Baroni, 2018). Lake and Baroni (2018) introduce the SCAN dataset to help study compositional generalization of neural networks, which has received increasing interests (Russin et al., 2019b;

Dessì and Baroni, 2019; Li et al., 2019c; Lake, 2019; Andreas, 2020; Gordon et al., 2020). Various benchmarks have been proposed including in the area of visual reasoning (Johnson et al., 2017b; Hudson and Manning, 2019), mathematics (Saxton et al., 2019), and semantic parsing (CFQ) (Keysers et al., 2020). However, no benchmark has been dedicated to machine translation in practice. We fill this gap by introducing a dataset with 216,000 instances and an average sentence length of 9.7 tokens.

3 Problem Definition

Following Keysers et al. (2020), *compositional generalization* is defined as the capacity to systematically generalize to novel combinations of components which are learned sufficiently during training. Key elements to measure compositional generalization include *atoms* and *compounds*. Specifically, **atoms** are primitive elements in the train set whereas **compounds** are obtained by composing these atoms. The research question is whether neural models perform well on unseen compounds. Take Table 2 for example, in the SCAN dataset, the atoms are simple commands such as "*jump*" and the composite command "*jump twice*" is a compound. In the CFQ, the compounds are questions such as "*Who directed Elysium*", and the atoms correspond to the primitive elements in the questions such as the predicate "*directed*", the question patterns "*Who [predicate] [entity]*" and the entities "*Elysium*".

In theory, compounds in MT can be defined as phrases, sentences or even document. In practice, however, we want to control the number of atoms in a novel compound for quantitative evaluation. In addition, it can be highly difficult to construct a large-scale dataset where novel compounds are sentences of practical sizes (the number of synthesized sentences increases exponentially with their length) while ensuring their grammatical correctness. Therefore, we constrain *compounds* to syntactic constituents, and define *atoms* as basic semantic components in constituents according to syntactic and semantic rules for forming constituents (Partee, 1995). As a result, we randomly assign multiple sentential contexts for investigating each novel compound. Table 2 shows a contrast between our dataset and existing datasets for compositional generalization in semantics.

Mistakes caused by weakness in computational generalization can be easily found in state-of-the-art NMT models. In particular, we train a Transformer-based model (Vaswani et al., 2017) on WMT17 En-Zh Dataset [1]. One sentence in the standard test set, "*but the problem is , with the arrival of durant , thompson 's* ***appearance rate*** *will surely decline , which is bound to affect his play*", is translated into "但问题是, 随着杜兰特的到来, 汤普森的外表肯定会下降, 这一定会影响到他的表演" (English: *but the problem is , with the arrival of durant , thompson 's will surely* ***look worse*** *, which is bound to affect his play*). The novel compound "*appearance rate*" is composed of two atoms (i.e., "*appearance*" and "*rate*"), both with a high frequency of more than 27,000 times in the training set. However, the sentence semantics is completely distorted due to the failure of semantic composition, which is possibly influenced by the context word "*play*". More importantly, as the overall translation highly overlaps with the reference, the model achieves a high score in similarity-based metrics such as BLEU, demonstrating that fatal translation errors can be overlooked under traditional evaluation metrics.

4 Dataset

Figure 1 gives an overview of our data construction process. We first source monolingual data (Section 4.1), and then build parallel data based by translation (Section 4.2). Then we synthesize a test set of novel compounds (Section 4.3), and offer an automatic evaluation method (Section 4.4).

4.1 Monolingual Data Source

Our goal is to focus on compositional generalization and minimize the influence of additional factors such as polysemy (Berard et al., 2019), misalignment (Munteanu and Marcu, 2005), and stylistic problems (Hovy et al., 2020). The dataset should ideally have following characteristics. First, the vocabulary size should be small and contain only words of high-frequency in order to avoid problems caused by rare words. In other words, variety of composition should come from combining different frequent words instead of word diversity, as suggested in (Keysers et al., 2020). Metaphorical words, which can increase the translation difficulty, should be excluded. Second, source sentences should not be too long or have complex syntactic structures. As a result, a sentence can be

[1] http://www.statmt.org/wmt17/

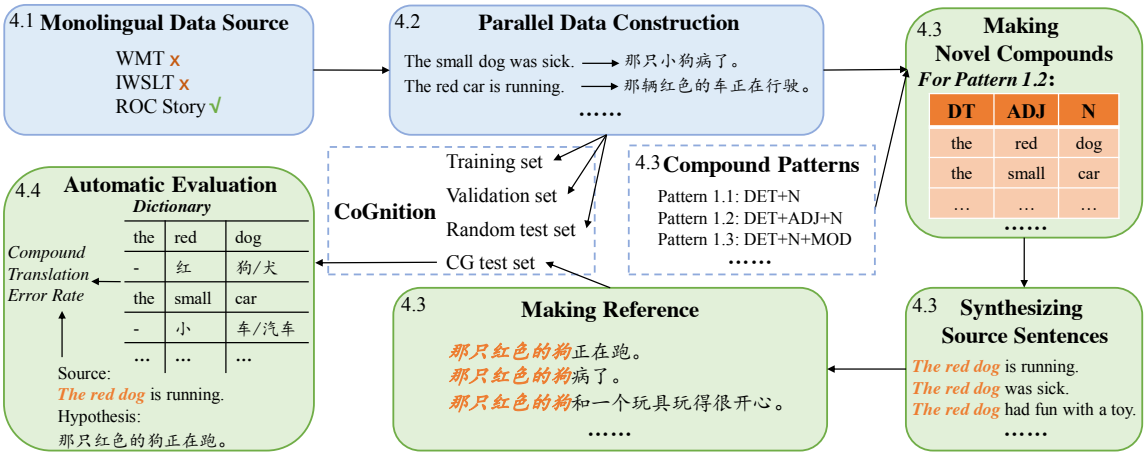

Figure 1: Summary of dataset construction.

Pattern #	Composition	Example
Pattern 1.1	DET+N	all the sudden **the waiter** screamed in pain .
Pattern 1.2	DET+ADJ+N	one day **another lazy lawyer** snapped and broke every window in the car .
Pattern 1.3	DET+N+MOD	**each doctor he liked** was talking to a friend on the phone .
Pattern 1.4	DET+ADJ+N+MOD	**every smart lawyer at the store** decided to go back next week .
Pattern 2.1	V+DET+N	she said she **liked the building** !
Pattern 2.2	V+DET+ADJ+N	he soon **met the special girl** named taylor .
Pattern 2.3	V+DET+N+MOD	she **took the child he liked** out to enjoy the snow .
Pattern 2.4	V+DET+ADJ+N+MOD	when taylor **saw the dirty car he liked** , he was amazed .
Pattern 3.1	P+DET+N	taylor felt really awful **about the bee** .
Pattern 3.2	P+DET+ADJ+N	**inside the small apartment** were some of my old toys .
Pattern 3.3	P+DET+N+MOD	taylor forgot **about the chair on the floor** !
Pattern 3.4	P+DET+ADJ+N+MOD	he jumped from the bench **towards the large airplane on the floor** .

Table 3: Compound patterns in the CG test set. Compounds are in bold and shown in sentence context.

translated literally, directly, and without rhetoric. Third, the corpus size should be large enough for training an NMT model sufficiently.

Widely-adopted corpora such as parallel data released on WMT and IWSLT[2] have large vocabularies and also contain noisy sentences and rich morphology (Li et al., 2019a), which do not fully meet our goal. We choose Story Cloze Test and ROCStories Corpora (Mostafazadeh et al., 2016, 2017) as our data source. The dataset is created for commonsense story understanding and generation, and consists of 101903 *5-sentence* stories. These stories are rather simple in items of vocabulary and syntax, but still contain rich phrases. In addition, the topic is constrained to daily life.

Since the vocabulary size of $42,458$ is large, we select the top $2,000$ frequent words as our vocabulary and extract sentences where the words are exclusively from the restricted vocab. Moreover, sentences that are longer than 20 words are removed. In this way, we finally obtain $216,246$ sentences for parallel data construction. More detailed statistics including comparison to WMT and IWSLT data are shown in Appendix B.

4.2 Parallel Data Construction

We take an MT post-editing method to construct parallel data, first using a public translation engine to obtain model-generated translations, and then requesting expert translators to post-edit them. The following aspects are highlighted:

- Ensure the fluency of translations.
- Ensure word-level matching between translated sentences and source sentences. Typically, every word should be correctly translated, without omission for legibility.

Finally, we obtain a parallel dataset of $216,246$ sentences in **CoGnition**, and randomly split it into three subsets: $196,246$ sentence pairs for **training**, $10,000$ sentence pairs for **validation**, and $10,000$ sentence pairs as the **random test set**. In addition

[2]https://wit3.fbk.eu/

to the above split, we additionally make a **compositional generalization test set**, which is described in the next section.

4.3 Compositional Generalization Test Set

We manually construct a special test set dedicated for evaluation of compositional generalization, by synthesizing new source sentences based on novel compounds and known contexts.

Designing Compound Patterns We use Berkeley Parser to obtain constituent trees (Kitaev and Klein, 2018). In CoGnition, noun phrases (NP), verb phrases (VP) and positional phrases (PP) are three most frequent constituents, accounting for 85.1% of all constituents, and thus we construct compounds based on them. According to syntactic and semantic rules (Partee, 1995), we choose basic semantic components as our atoms including determiners (DET), nouns (N), verbs (V), prepositions (P), adjectives (ADJ), and postpositive modifiers (MOD). Specifically, postpositive modifiers include prepositional phrases and relative clauses, and can contain multiple words. We consider them as a single atom due to their semantic inseparability. In this way, we generate 4 compound patterns for NP, VP, and PP, respectively, which are listed in Table 3 with corresponding examples.

Making Novel Compounds We use Stanza (Qi et al., 2020) to obtain POS tagging for each word in training sentences. We construct novel compounds by first selecting atom candidates with relatively consistent translation in the training set. The frequency of candidate atoms covers a wide range from 34 to 73518. We list full set of atom candidates in Table 4. For constructing compounds, we enumerate all possible combinations of atoms according to the patterns in Table 3, and then remove those that are ungrammatical or likely to cause ethic issues, obtaining 2,160 compounds finally. We do not deliberately make all compounds unseen, yet only 0.93% of them appear in the training data.

Synthesizing Source Sentences We embed the compounds in specific context to form complete source sentences. Concretely, we first apply Berkeley Parser on the training sentences to obtain sentence templates, where certain constituents are replaced by placeholders according to their constituent types, e.g., "*NP-placeholder spent a lot of time to set up a wedding* .". Then we select 5 sentence templates for each constructed compound accordingly, so that every compound can be evaluated under 5 different contexts. To distinguish from VP and PP, we put NP compounds only in sentences with the placeholder outside VP and PP.

Making Reference To maintain statistical consistency, target translations of synthetic sentences are also obtained using the same MT post-edit approach. In addition to the annotation principles listed in 4.2, we set several additional rules:

- Filter sentences with ethical issues and replace them with other synthetic ones.
- Ensure the accuracy of compound translation.

Finally, we obtain a compositional generalization test set (**CG test set**) of $10,800$ parallel sentences. The final dataset statistics is shown in table 5.

4.4 Automatic Evaluation

We mainly adopt human evaluation for the experiments of this paper (Section 5) for ensuring reliability of findings. Despite its accuracy, human evaluation can be expensive. To facilitate fast evaluation in future research, we introduce an automatic evaluation approach to quantify a model's generalization ability on our CG test set.

In particular, we manually construct a dictionary for all the atoms based on the training set (See Appendix C). The prerequisite of correctly translating one compound is that all of the atom translations should be contained. Besides, in most cases the translation of nouns should be placed after that of other atoms. Based on this, we design a heuristic algorithm to determine whether compounds are translated correctly. With the human annotation as ground truth, our automatic evaluation tool achieves a precision of 94.80% and a recall of 87.05%, demonstrating it can serve as an approximate alternative to human evaluation.

5 Experiments

We conduct experiments on CoGnition dataset and perform human evaluation on the model results.

5.1 Settings

We tokenize the English side using Moses tokenizer and do not apply byte pair encoding (BPE) (Sennrich et al., 2016) due to the small vocabulary (i.e., 2000). The Chinese sentences are segmented by

Type	Candidates
DET	the, every, any, another, each
N	car, dog, girl, doctor, boyfriend, apartment, child, sandwich chair, farm, building, hat, waiter, airplane, lawyer, peanut, farmer, clown, bee
ADJ	small, large, red, special, quiet, empty, dirty, lazy, smart, fake, silly
MOD	he liked, at the store, on the floor
V	took, told, found, asked, saw, left, gave, lost, liked woke, stopped, invited, met, caught, heard, hated, watched, visited, chose
P	to, for, on, with, from, about, before, like, around inside, without, behind, under, near, towards, except, toward

Table 4: Atoms used in constructing compounds, sorted by frequency in the training set.

Split	# Samples
Training set	196,246
Validation set	10,000
Random test set	10,000
CG test set	10,800

Table 5: Statistics of CoGnition Dataset.

jieba segmenter[3]. We employ BPE with 3,000 merge operations, generating a vocabulary of 5,500 subwords.

We focus on Transformer (Vaswani et al., 2017) because of its state-of-the-art performance on machine translation (Edunov et al., 2018b; Takase and Kiyono, 2021; Raffel et al., 2020; Zhu et al., 2020; Liu et al., 2020b) and better performance on existing compositional generalization dataset (Daniel et al., 2019). We implement our model using BASE configuration provided by Fairseq (Ott et al., 2019). The model consists of a 6-layer encoder and a 6-layer decoder with the hidden size 512. We tie input and output embeddings on the target side. The model parameters are optimized by Adam (Kingma and Ba, 2015), with $\beta_1 = 0.1$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. The model is trained for 100,000 steps and we choose the best checkpoint on validation set for evaluation.

We report character-level BLEU scores using SacreBLEU (Post, 2018) to measure the overall translation performance. In addition, we request expert translators to annotate the correctness of compound translation. Translators are asked to only focus on examining whether the compound itself is translated correctly or not, disregarding errors in context. Specifically, a compound is correct only if its translation contains semantic meaning of all atoms and is fluent in human language. Since each of the 2,160 compounds is provided with 5 contexts, we can compute the translation error-rate for each compound.

[3]https://github.com/fxsjy/jieba

5.2 Main Results

Table 6 shows the results. Besides the *CG test set*, we also list results on three of its subsets, which only contain NP, VP or PP compounds respectively. The model achieves a 69.58 BLEU score on the *random test set*, which partly indicates distributional consistency and quality of the dataset. In comparison, the performance on the *CG test set* drops dramatically by more than 20 BLEU points. Given that the only difference between synthetic sentences and training sentences is the unseen compounds (i.e., contexts are seen in training), the decrease of 20 BLEU points indicates that unseen compounds pose a significant challenge, which is however easy to be overlooked in traditional evaluation metrics. For example, the model mis-translates "*alas , he became sick from eating all of the peanut butter* ***on the ball***" into "唉，他因为吃掉了球场上所有的花生酱而生病了" (English: *alas , he became sick from eating all of the peanut butter* ***on the field***). With a minor mistake on the compound "*on the ball*", the model achieves a sentence-level BLEU of 61.4, despite that the full sentence meaning is largely affected. In other words, the BLEU score of 69.58 can be misleading since novel compounds can be rare in the *random test set*. Such mistakes in generalizing new compounds can severely hinder overall performance of translation engines in practice, as shown earlier in Table 1. Also, we calculate BLEU for the original *training* sentences that provide contexts for the CG test set (row 3). The model achieves 99.74 BLEU, further demonstrating that the performance degradation is mainly caused by the unseen compounds.

Instance-wise, 27.31% compounds are translated incorrectly. However, when aggregating all 5 contexts, 61.62% compounds suffer at least one incorrect translation. This suggests that a well-trained NMT model is not robust in translating compounds, though all atoms within them are highly frequent in

Test Set	Error Rate		BLEU
	Instance	Aggregate	
Random-test	-	-	69.58
Train	-	-	99.74
CG-test	27.31%	61.62%	48.66
CG-test/NP	21.94%	54.03%	51.29
CG-test/VP	22.25%	55.56%	47.55
CG-test/PP	37.72%	75.28%	47.14

Table 6: BLEU score and compound translation error rate on the *random test* set and the *CG test set*.

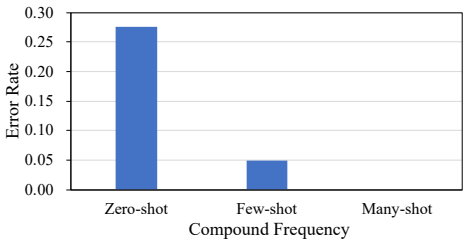

Figure 2: Effect of compound frequency on compound translation error rate.

the training set. We also observe that the error rate of PP compounds, 37.72%, is much higher than the other two, 21.94% and 22.25%, which we will discuss in detail in the following section.

6 Analysis

We conduct experiments to explore in what situations the model is error-prone by considering compound frequency, compound length, compound structure, atom frequency, atom co-occurrence, and the complexity of external context.

6.1 Compound Frequency

Intuitively, compounds with higher frequencies in the training set are easier to infer. We classify compounds according to their frequency levels, including many-shots (frequency higher than 10), few-shots (frequency from 1 to 10) and zero-shot, and show the error rate for each bucket in Figure 2. The model translates all the many-shots compounds correctly. For few-shot compounds, translation error rate increases to 5.00%, but is still much lower than zero-shot compounds with an error rate of 27.53%. The result suggests the model is good at memorizing correspondence between sentence segments. However, the model deteriorates severely when test samples are unseen in the training set, which further confirms model's weakness in compositional generalization (Lake and Baroni, 2018).

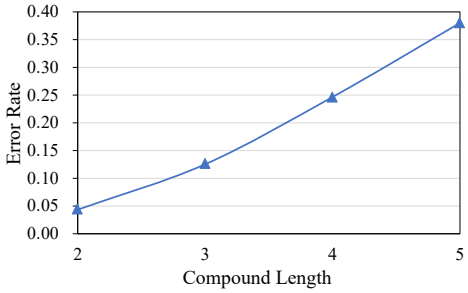

Figure 3: Effect of compound length on compound translation error rate.

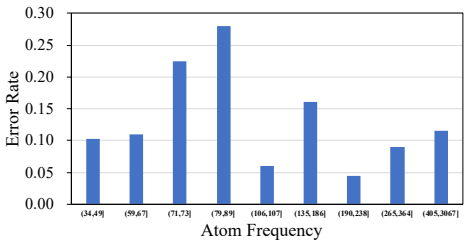

Figure 4: Effect of atom frequency on compound translation error rate.

6.2 Compound Length

As shown in Figure 3, the error rate grows with the increase of compound length (i.e., the number of atoms in a compound). Only 4.50% of the shortest compounds are translated incorrectly, each of which consists of a determiner and a noun. The error rate increases to 13.72% when the compound length grows to 3 atoms (e.g., "*the smart lawyer*"). The longest compounds contain a determiner, a noun, an adjective, a modifier and a preposition or verb in each of them, e.g., "*taking every special chair he liked*". The error rate increases to 36.63%, demonstrating that it is more difficult to generalize in longer compounds, which contain richer semantic information. We conjecture that if the range of *compound* is further expanded, the error rate will be much higher.

6.3 Atom Frequency

We empirically divide compounds into multiple groups according to the minimum frequency of their atoms, where each group consists of similar numbers of compounds. The intuition is that the atom with low frequency might be difficult to translate and therefore hinders the whole compound translation. We fix the compound length to 3 in order to reduce effects of compound length.

As shown in Figure 4, the error rate has no strong

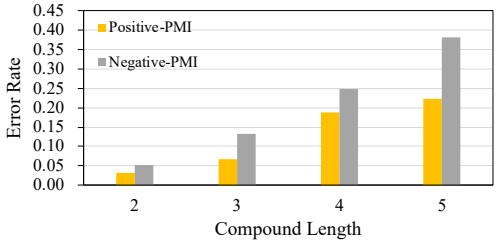

Figure 5: Effect of atom co-occurrence on compound translation error rate.

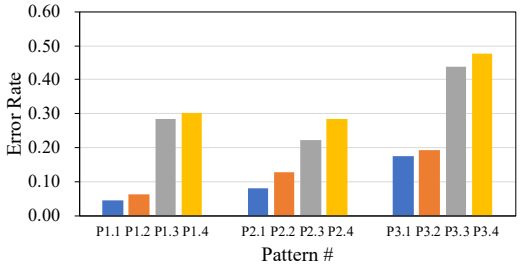

Figure 6: Compound translation error rates of different patterns.

correlation with the atom frequency. This can be because all atoms in our corpus are simple and relatively frequent and thus it is easy for the NMT model to memorize the semantics of most atoms. Therefore, simply increasing atom frequency does not enhance model's generalization ability of novel compounds. We observe similar patterns for compounds of other lengths (Appendix A).

6.4 Atom Co-occurrence

Although the NMT model may never see a compound, there can exist many local segments where atoms co-occur. For example, in the unseen compound "*the smart lawyer*", "*smart*" and "*lawyer*" may occur within some training sentences. Intuitively, the compounds of which atoms co-occur more frequently may be translated better. We calculate pointwise mutual information (PMI) and compare error rates of compounds with positive or negative mean PMI scores (MPMI):

$$\text{MPMI}(C) = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \text{PMI}(a_i, a_j), \quad (1)$$

where a_i is the i-th atom in the compound C, N is the compound length, M is the number of possible combinations of two atoms, and PMI score is computed as:

$$PMI(x, y) = log \frac{p(a_i, a_j)}{p(a_i)p(a_j)}, \quad (2)$$

where the probabilities $p(a_i)$ and $p(a_i, a_j)$ are obtained by dividing the number of n-grams in which one word or both words occur by the total number of n-grams[4].

We divide compounds into 4 groups by their length and compare error rates within each group. As shown in Figure 5, across all groups, the error rates with positive mean PMI scores are lower than those with negative ones, verifying our hypotheses.

[4]We use 5-gram here

6.5 Linguistic Factors

Figure 6 shows the error rates of all compound patterns in Table 3. The MOD atom exerts salient influence on translation error rate. The error rate of compounds with MOD is 19.78% higher than those without on average. In contrast, adding ADJ into compounds only increases error rate by 2.66%. The major difficulty caused by MOD is word reordering. One can translate "*the small dog*" monotonically without adjusting word order. However, compounds like "*the dog he liked*" require the model to recognize "*he liked*" as MOD and put its translation before that of "*the dog*" in Chinese. We find many cases where the model translates such compounds without reordering or breaking the connection between nouns and modifiers.

Across these groups, we can see that the error rate of NP (Pattern 1.*) is generally lower than that of VP (Pattern 2.*) and PP (Pattern 3.*). Such phenomenon is more obvious for the patterns without MOD. The reason is that compounds in Pattern 1.* are generally shorter and contain less semantic and syntactic information. However, the error rates of Pattern 2.3 and 2.4 are lower than other patterns with MOD (i.e., Pattern 1.3, 1.4, 3.3 and 3.4), indicating the model performs better in "*V+DET(+ADJ)+NN+MOD*". This can be because under certain situations the MOD can be useful for correctly translating verbs, which are more commonly seen in the training set, e.g., "***found** the chair **on the floor***".

We also observe that compounds of PP (Pattern 3.*) are more difficult to translate compared with VP (Pattern 2.*), although both types of compounds share the same compound length. In the training set, verbs typically have consistent translations, whereas the meanings of prepositions vary with contexts. Therefore prepositional compounds are more difficult to translate as more context infor-

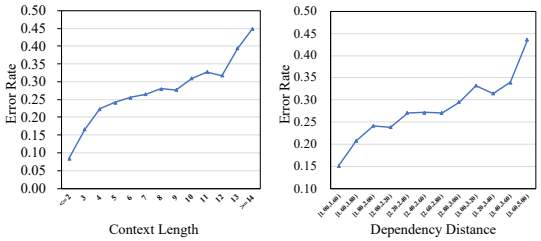

Figure 7: Effect of external context on compound translation error rate.

mation is required to ground their meanings.

6.6 Effect of External Context

Due to the nature of NMT, the semantic representation of each compound is context-aware. Intuitively, translation of compounds is also influenced by external context, which is sentential in our case but can also be document-level in practice. We investigate effects of context lengths and sentence comprehension difficulty. In particular, the context length is calculated by subtracting the sentence length by the number of words in the compound. Comprehension difficulty of the training sentences which provide contexts, is quantified by the dependency distance (Liu, 2008): $\mathrm{MMD}(x) = \frac{1}{N-1}\sum_{i}^{N} D_i$, where N is the number of words in the sentence and D_i is the dependency distance of the i-th syntactic link of the sentence.

The results are shown in Figure 7. The translation error rate increases stably with the context length as well as the dependency distance. These observations demonstrate that the generalization for novel compounds correlates strongly with context complexity. Sentences with higher dependency distances are harder for model to comprehend during training. Given that our test sentences are restricted to 20 words, compositional generalization can be more challenging in practice where average sentence lengths can be much longer.

7 Conclusion

We proposed a dedicated parallel dataset for measuring compositional generalization of NMT and quantitatively analyzed a Transformer-based NMT model manually. Results show that the model exhibits poor performance on novel compound translation, which demonstrates that the NMT model suffers from fragile compositionality, and it can be easily overlooked under transitional metrics. To the best of our knowledge, we are the first one to propose a practical benchmark for compositionality of NMT, which can be a testbed for models tailored for this specific problem.

8 Ethics Consideration

As mentioned, we collected our data from Story Cloze Test and ROCStories Corpora that all are public to academic use, and they contain no sensitive information (Mostafazadeh et al., 2016, 2017). The legal advisor of our institute confirms that the sources of our data are freely accessible online without copyright constraint to academic use. Our data construction involves manual annotation. Annotators were asked to post-edit machine translation and filter out samples that may cause ethic issues, which do not involve any personal sensitive information.

We hired 4 annotators who have degrees in English Linguistics or Applied Linguistics. Before formal annotation, annotators were asked to annotate 100 samples randomly extracted from the dataset, and based on average annotation time we set a fair salary (i.e., 32 dollars per hour) for them. During their training annotation process, they were paid as well.

Acknowledgment

Yue Zhang is the corresponding author. We thank all reviewers for their insightful comments. This work is supported by National Natural Science Foundation of China (NSFC) under grant No.61976180 and a grant from Lan-bridge Information Technology Co., Ltd. We thank colleagues from Lan-bridge for examining data and evaluating results. Major contributors include Xianchao Zhu, Guohui Chen, Jing Yang, Jing Li, Feng Chen, Jun Deng and Jiaxiang Xiang.

References

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.

Jasmijn 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.

Yonatan Belinkov and Yonatan Bisk. 2018a. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yonatan Belinkov and Yonatan Bisk. 2018b. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Alexandre Berard, Ioan Calapodescu, Marc Dymetman, Claude Roux, Jean-Luc Meunier, and Vassilina Nikoulina. 2019. Machine translation of restaurant reviews: New corpus for domain adaptation and robustness. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 168–176, Hong Kong. Association for Computational Linguistics.

Arianna Bisazza and Clara Tump. 2018. The lazy encoder: A fine-grained analysis of the role of morphology in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2871–2876, Brussels, Belgium. 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*.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333, Florence, Italy. Association for Computational Linguistics.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1766, Melbourne, Australia. Association for Computational Linguistics.

Noam Chomsky. *Syntactic Structures*. Mouton, The Hague.

Jeanne E. Daniel, Willie Brink, Ryan Eloff, and Charles Copley. 2019. Towards automating healthcare question answering in a noisy multilingual low-resource setting. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 948–953, Florence, Italy. Association for Computational Linguistics.

Roberto Dessì and Marco Baroni. 2019. CNNs found to jump around more skillfully than RNNs: Compositional generalization in seq2seq convolutional networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3919–3923, Florence, Italy. Association for Computational Linguistics.

Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1150–1159, Vancouver, Canada. Association for Computational Linguistics.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018. On adversarial examples for character-level neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 653–663, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018a. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018b. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Marzieh Fadaee and Christof Monz. 2020a. The unreasonable volatility of neural machine translation models. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 88–96, Online. Association for Computational Linguistics.

Marzieh Fadaee and Christof Monz. 2020b. The unreasonable volatility of neural machine translation models. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 88–96, Online. Association for Computational Linguistics.

Jonathan Gordon, David Lopez-Paz, Marco Baroni, and Diane Bouchacourt. 2020. Permutation equivariant models for compositional generalization in language. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Shilin He, Zhaopeng Tu, Xing Wang, Longyue Wang, Michael Lyu, and Shuming Shi. 2019. Towards understanding neural machine translation with word importance. 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 953–962, Hong Kong, China. Association for Computational Linguistics.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. Can you translate that into man? commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, WA, USA. Association for Computational Linguistics*.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017a. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017b. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society.

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 *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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*.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Brenden M. Lake. 2019. Compositional generalization through meta sequence-to-sequence learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9788–9798.

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.

Bei Li, Yinqiao Li, Chen Xu, Ye Lin, Jiqiang Liu, Hui Liu, Ziyang Wang, Yuhao Zhang, Nuo Xu, Zeyang Wang, Kai Feng, Hexuan Chen, Tengbo Liu, Yanyang Li, Qiang Wang, Tong Xiao, and Jingbo Zhu. 2019a. The NiuTrans machine translation systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 257–266, Florence, Italy. Association for Computational Linguistics.

Yuanpeng Li, Liang Zhao, Jianyu Wang, and Joel Hestness. 2019b. 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.

Yuanpeng Li, Liang Zhao, Jianyu Wang, and Joel Hestness. 2019c. 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.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Xiaodong Liu, Kevin Duh, Liyuan Liu, and Jianfeng Gao. 2020a. Very deep transformers for neural machine translation. *CoRR*, abs/2008.07772.

Xiaodong Liu, Kevin Duh, Liyuan Liu, and Jianfeng Gao. 2020b. Very deep transformers for neural machine translation. *CoRR*, abs/2008.07772.

João 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.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14014–14024.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Richard Montague. Universal grammar. In Richmond H. Thomason, editor, *Formal Philosophy: Selected Papers of Richard Montague*, 222–247. Yale University Press, New Haven, London.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. LSDSem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, Valencia, Spain. Association for Computational Linguistics.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Barbara Partee. 1995. Lexical semantics and compositionality. *An invitation to cognitive science: Language*, 1:311–360.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Vikas Raunak, Vaibhav Kumar, and Florian Metze. 2019. On compositionality in neural machine translation. *CoRR*, abs/1911.01497.

Jake Russin, Jason Jo, Randall C. O'Reilly, and Yoshua Bengio. 2019a. Compositional generalization in a deep seq2seq model by separating syntax and semantics. *CoRR*, abs/1904.09708.

Jake Russin, Jason Jo, Randall C. O'Reilly, and Yoshua Bengio. 2019b. Compositional generalization in a deep seq2seq model by separating syntax and semantics. *CoRR*, abs/1904.09708.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–

1534, Austin, Texas. Association for Computational Linguistics.

Sho Takase and Shun Kiyono. 2021. Lessons on parameter sharing across layers in transformers. *CoRR*, abs/2104.06022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. 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 4396–4406, Hong Kong, China. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019b. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3070–3079, Online. Association for Computational Linguistics.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating BERT into neural machine translation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Atom Frequency

For compounds of other lengths, we also compute their error rates with respect to minimum atom frequency. As shown in Figure 8, 9 and 10, the error rate does not correlates with atom frequency across all compound lengths.

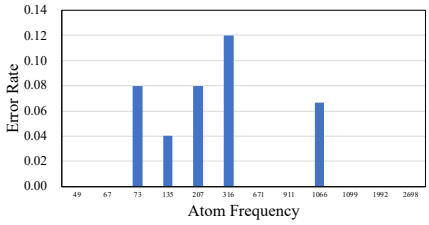

Figure 8: Effect of atom frequency with compound length fixed to 2.

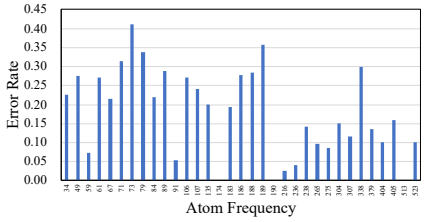

Figure 9: Effect of atom frequency with compound length fixed to 4.

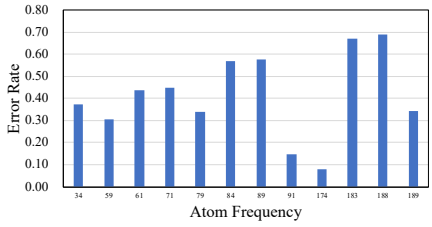

Figure 10: Effect of atom frequency with compound length fixed to 5.

B Data Statistics

Table 7 and Table 8 lists statistics of several monolingual data sources, compared with the data source (ROC-Filter) used in constructing the CoGnition dataset. We can see that our dataset has both shorter sentences and vocabulary made up of more frequent words.

Data \ Property	Vocab	#Tokens	#Sents
WMT17 En-Zh	1,201,752	518,286,577	20,616,495
IWSLT17 En-Zh	70,950	4,715,201	231,266
ROC-Original	42,458	5,283,521	532,093
ROC-Filter	2,000	2,096,524	216,246

Table 7: Statistics of data sources: vocabulary size, number of tokens and number of sentences.

Data \ Property	Avg Len	Avg Freq	Min Freq
WMT17 En-Zh	25.1	431.3	1
IWSLT17 En-Zh	20.4	66.5	1
ROC-Original	9.3	124.4	1
ROC-Filter	9.7	1048.3	35

Table 8: Statistics of data sources: average sentence length, average token frequency and minimum token frequency.

C Lexicon

Part of the lexicon for automatic evaluation is shown in Table 9.

Atom	Lexical Translation
dog	狗/犬
doctor	医生
sandwich	三明治
hat	帽
waiter	服务员
lawyer	律师
peanut	花生
farmer	农夫/农场主/农贸市场/农民
small	小
red	红
dirty	脏
lazy	懒
smart	聪明/明智/智能
the	-
every	每/所有
any	任何
another	另/又/再/还/别
each	每
he liked	他喜欢的
at the store	店里/商店

Table 9: Lexicon for automatic evaluation.