Grammar-based Machine Translation of a Dataset for Measuring Compositional Generalization in Semantic Parsing

Zi Wang
zrw348@alumni.ku.dk

1 Introduction

Compositional generalization refers to the ability exhibited by human intelligence to understand novel combinations of familiar units like words. With such ability, people are naturally able to understand the compositions of known components, potentially to an "infinite" extent (Chomsky, 1965). While compositional generalization is generally demonstrated by humans in various domains (Johnson et al., 2017; Higgins et al., 2018), it has been extracted as a research focus and is predominately measured in natural language. However, research on state-of-the-art neural networks applied to semantic parsing has suggested that they fail to generalize compositionally akin to humans (Lake and Baroni, 2018; Bastings et al., 2018).

Semantic parsers enable machines to understand natural language utterances, typically by parsing them as logical formalism. A growing amount of research has been investigating the compositional generalization ability of semantic parsers based on elaborated datasets, typically synthetic corpora(eg. CFQ (Keysers et al., 2019)) automatically generated according to curated rules (Lake and Baroni, 2018; Kim and Linzen, 2020). To further extend this topic to a multilingual category MCWQ (Cui et al., 2022) was created based on CFQ through a proposed methodology mainly including knowledgebase migration and question translation. While Google Translate (Wu et al., 2016), a neural-based model trained on large-scale corpora is adopted in creating MCWQ, we argue that meaning preservation during translation is vulnerable in this methodology especially considering the synthetic nature of compositional corpora. It is mainly due to the unstable mapping from source to target languages in neural-based translation models, at both the lexical level and the syntactic structural level, i.e., potentially generating different mappings for even the same terms or sentence patterns, which conflicts with the parallel principle in compositionality among multilingual corpora.

While neural machine translation applies to more general scenarios, in this project, we propose to utilize rule-based machine translation to extend MCWQ corpora with parallel versions, since the rule-based model constructs stable mappings strictly following manual rules thus avoiding unexpected variations in lexicon and structure. For this purpose, we build a SCFG-based machine translation framework involving functional components (for parsing, disambiguation, etc.) based on open-source toolkit. As an instance, we build the grammar rules from English to Japanese to create the Japanese branch of the corpora. Furthermore, the implemented system successfully obtains the expected translation results in Japanese, a language with a substantial typological differences from English, indicating the feasibility of the rule-based methodology in compositional corpora extension. We believe our work gives a more reliable cross-lingual benchmark by minimizing the confound of translation quality (See Section 6.2). In addition, benefiting from the resulting expectable sentence patterns, finer statistical analysis (eg. compound divergence) are applicable to our branch. [1]

2 Preliminaries

2.1 Definition of Formal Grammar

Here we give formal definitions for context-free grammar and synchronous context-free grammar which are the basis of our methodology. The definition below largely refers to Williams et al. (2016).

Context-Free Grammar A *context-free grammar (CFG)* G is defined by a 4-tuple $G = \langle T, N, S^\dagger, R \rangle$, where T is a finite set of *terminal* symbols; N is a finite set of *non-terminal* symbols; $S^\dagger$ is a *start* symbol belonging to N; R is a set of

[1] Our code and generated dataset are public at https://github.com/ziwang-klvk/CFQ-RBMT

relation in $N \times (N \cup T)^*$ called *production rules* in which each item has the form $A \rightarrow \alpha$, where A is *non-terminal* and α is a string of terminals and non-terminals.

Synchronous Context-Free Grammar A *synchronous context-free grammar* is defined by a 4-tuple $G_{sync} = \langle T, N, S^{\dagger}, R_{sync} \rangle$, where T, N, $S^{\dagger}$ are as defined for CFG; R_{sync} is a set of *synchronous rules* in which each item is a pair of CFG *production rules* represented as $A \rightarrow \alpha$ and $B \rightarrow \beta$, where A and B are non-terminals and α and β strings of terminals and non-terminals. The paired rules are under constraints that α and β involve the same number of non-terminals, and each non-terminal in α is *exclusively* paired with one non-terminal in β.

3 Background

Datasets for Evaluating Compositional Generalization Much previous work on compositional generalization investigated how to measure the compositional ability of semantic parsers, conventionally within the English language. Lake and Baroni (2018) proposed to evaluate the sequence-to-sequence models on the task of interpreting formalized natural language commands into ordered actions, namely SCAN, which provided evidence that the systematicity distribution gap between training and test sets is fundamental for measuring compositional capacity. Keysers et al. (2019) brought this task to a more realistic scenario of knowledge base question answering and further determined the distribution gap as compound divergence, based on which CFQ was synthetically generated and split, resulting in 3 Maximum Compound Divergence (MCD) splits. Similarly, Kim and Linzen (2020) created COGS in this synthetic fashion, whereas it consists of less confined sentence patterns and its generative rules follow a stronger definition of training-test distribution gap. While the mentioned previous work differs in data expressions and splitting strategy, rule-based approaches are commonly adopted for dataset generation; as Kim and Linzen (2020) put it, such approaches allow maintaining "full control over the distribution of inputs" , the crucial factor for valid compositionality measurement.

Goodwin et al. (2022) benchmarked dependency parsing in compositional generalization by introducing gold dependency parses for CFQ dataset. For this purpose, a full coverage context-free grammar over CFQ was constructed benefiting from the synthetic nature of the dataset.

Cui et al. (2022) first benchmarked compositional generalization over multilingual data with Multilingual Compositional Wikidata Questions (MCWQ), which is the basis of our work. MCWQ comprises the Hebrew, Chinese, Kannada and English versions of questions from CFQ and corresponding queries linked to Wikidata [2] migrated from Freebase. As stated in section 1, the MCWQ questions were translated with Google Cloud Translation, which disregards the "control over distribution" and is inconsistent with previous work. Thus the main aspect where our work differs from MCWQ is that the Japanese branch is created following strict grammar rules to guarantee controllability during translation; such controllability ensures that the translations as well as their composing corpus is determinable and analyzable in terms of both specific atoms/compounds and their overall distribution, thus enabling us to create the dataset for the purpose of measuring compositional generalization systematically.

Machine Translation The research on machine translation commenced almost as early as the birth of the computer; during the decades of development, various methodologies and technologies were introduced to this topic. To informally categorize the most popular models, we can divide them into pre-neural models and neural-based models. Pre-neural MT (Wu, 1996; Marcu and Wong, 2002; Koehn et al., 2003; Chiang, 2005) typically includes manipulation of syntax and phrases, whereas the neural-based MT (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Vaswani et al., 2017) refer to those employing neural networks such as the state-of-the-art sequence-to-sequence models. However, oriented to general applications, most models in existence commonly rely on learned statistical prior, even for the pre-neural models.

The desiderata in our work intuitively exclude the methods with the necessity of large-scale training. The most relevant works to ours were by Wu (1996, 1997) who first applied variants of SCFG with distinctions mainly in formalism to MT (Chiang, 2006). The SCFG is a generalization of CFG as generating coupled strings instead of single ones

[2] Wikidata is an open KB where each item is allocated a unique, persistent identifier (QID). https://www.wikidata.org/

(See Section 2.1); this property was exploited by pre-neural MT works to satisfy complex syntactic reordering during translation. In this work, we exclude the statistical component in previous work and manually build the SCFG transduction due to the synthetic nature of CFQ; we specifically call it "rule-based" instead of "syntax-based" to emphasize this subtle difference. In addition, a formally defined SCFG variant tends to be symmetrical regarding both languages (Wu, 1997), while we implement a simplified yet functionally identical version only for one-way transduction.

Cross-lingual Learning Cross-lingual learning has been increasingly researched recently, where popular technologies in NLP are generally adapted for representation learning over multiple languages (Kenton and Toutanova, 2019; Conneau et al., 2020; Xue et al., 2021). Meanwhile, transfer learning is widely leveraged to overcome the data scarcity of low-resource languages (Cui et al., 2019; Hsu et al., 2019). However, as modelling research is developed against corresponding benchmarks, part of them also revealed that cross-lingual dataset suffers "translation artifact" when created under arbitrary usage of general machine translation systems (Artetxe et al., 2020; Wintner, 2016). Longpre et al. (2021) thus proposed Multilingual Knowledge Questions and Answers (MKQA), a large-scale multilingual corpus (yet not for evaluating compositional generalization) avoiding this issue, yet through enormous human efforts. In contrast, Cui et al. (2022) adopted Google Translate to easily obtain parallel versions for CFQ questions while sacrificing meaning preservation. The desiderata for our work is partially a balance of these two methodologies.

The mentioned data scarcity is rather notable in cross-lingual semantic parsing. Whilst several studies have explored the zero-shot cross-lingual semantic parsing, the development is nevertheless relatively slow with limited annotated data for language other than English. Cui et al. (2022) created a 4-language parallel corpora MCWQ as a benchmark to address this gap, specially in knowledge base question answering and compositional generalization. Our work further fill this gap by giving more reliable benchmark and introduce the first Japanese branch.

4 Methodology

4.1 Toolkit

We base our method on the Universal Rule-Based Machine Translation toolkit (URBANS; Nguyen, 2021) and modified it to a framework supporting SCFG for practical use. The original URBANS translates text with manually configured source language grammar, source-to-target transduction rules and corresponding dictionary. Specifically, the toolkit parses the sentence as syntactic trees based on CFG (See Section 2.1) and reorders the whole sentence hierarchically based on the transduction rules. We give the definition of URBANS framework with example rules as follows:

SOURCE GRAMMAR

$$\begin{aligned} \text{S} &\rightarrow \text{NPQ VP} \\ \text{VP} &\rightarrow \text{V NP} \\ \text{V} &\rightarrow \text{directed} \\ \text{NPQ} &\rightarrow \text{who} \\ \text{NP} &\rightarrow \text{Inception} \end{aligned} \tag{1}$$

TRANSDUCTION RULE

$$\text{VP} \rightarrow \langle \text{V NP}, \text{NP V} \rangle \tag{2}$$

DICTIONARY

$$\begin{aligned} &\langle \text{who}, \overset{\text{dare}\ \text{ga}}{\text{誰が}} \rangle \\ &\langle \text{directed}, \overset{\text{kan}\ \text{toku}\ \text{shi}\ \text{ma}\ \text{shi}\ \text{ta}}{\text{監督しました}} \rangle \\ &\langle \text{Inception}, \text{Inception}\overset{\text{wo}}{\text{を}} \rangle \end{aligned} \tag{3}$$

This set of rules supported by URBANS can translate the English sentence *"who directed Inception"* into Japanese as "$\overset{\text{dare}\ \text{ga}}{\text{誰が}}\text{Inception}\overset{\text{wo}}{\text{を}}\overset{\text{kan}\ \text{toku}\ \text{shi}\ \text{ma}\ \text{shi}}{\text{監督しまし}}\overset{\text{ta}}{\text{た}}$", by simply referring to the TRANSDUCTION RULE and DICTIONARY after parsing with the SOURCE GRAMMAR. Notice that the transduction rule in this case indicates the different verb phrase (VP) and noun phrase (NP) order in English and Japanese.

Although URBANS provides interfaces of well-implemented syntactic-tree manipulation methods, due to its rigid utilization of the CFG module provided by the Natural Language Toolkit (NLTK; Bird et al., 2009), it is nevertheless a restricted version of one-way SCFG. For example, the rules in this framework are not able to handle polysemy and complex inflections such as "star", which potentially means *"an actor"* or *"having ... as an actor"*

depending on its part-of-speech (POS). Considering all the distinctions in formalism from SCFG as defined by Chiang (2006), the crucial element resulting in this functional gap is that the dictionary in URBANS framework lacks a link to the non-terminals, which interdicts the lexicons from translated based on the top-down syntactic information.

To enable URBANS to handle these realist problems, we adapted it as functionally identical with SCFG through simple yet effective adjustment to DICTIONARY, namely "tag dictionary". Specifically, the modified dictionary build the lexicon-level mappings constrained with the intermediate tag of the term. Consider the *"star"* example mentioned above:

TAG DICTIONARY

$$\begin{aligned} \text{NP} &\rightarrow \langle \text{star}, \overset{\text{shu}\ \text{en}}{\text{主演}} \rangle \\ \text{V} &\rightarrow \langle \text{star}, \overset{\text{shu}\ \text{en}\ \text{sa}\ \text{se}\ \text{ma}\ \text{su}}{\text{主演させます}} \rangle \\ \text{VPast} &\rightarrow \langle \text{starred}, \overset{\text{shu}\ \text{en}\ \text{sa}\ \text{se}\ \text{ma}\ \text{shi}\ \text{ta}}{\text{主演させました}} \rangle \\ \text{VPastPass} &\rightarrow \langle \text{starred}, \overset{\text{shu}\ \text{en}\ \text{sa}\ \text{se}\ \text{ra}\ \text{re}\ \text{ma}\ \text{shi}\ \text{ta}}{\text{主演させられました}} \rangle \end{aligned} \tag{4}$$

(4) demonstrates how this mechanism handles the polysemy and inflections in both source and target languages.

Apart from the crucial element modified as above, another notable formal difference between our grammar and SCFG is that the latter one assigns a *boxed number* to each right-side symbol in a production, for which non-terminal rules are defined as (5).

SCFG

$$\begin{aligned} \text{S} &\rightarrow \langle \text{NP}\boxed{1}\ \text{VP}\boxed{2},\ \text{NP}\boxed{1}\ \text{VP}\boxed{2} \rangle \\ \text{VP} &\rightarrow \langle \text{V}\boxed{3}\ \text{NP}\boxed{4},\ \text{NP}\boxed{4}\ \text{V}\boxed{3} \rangle \end{aligned} \tag{5}$$

In the formal definition of SCFG, the linking boxed numbers are explicit symbolic representation for the synchronous principle. In our practical usage, however, this linking mechanism is not functionally necessary since we can always rewrite the rules as well as rename the tags (the linking number can even be explicitly named in the tags where necessary).

With the demonstration above, we stress again that our implemented framework which comprises (1), (2) and (4) and originates from URBANS supports a functionally identical version of SCFG for the one-way translation purpose.

4.2 Transduction Grammar

We build the English-Japanese transduction grammar (one-way SCFG) within the implemented framework. The synthetic nature of CFQ indicates that it has limited sentence patterns and barely causes ambiguities; Goodwin et al. (2022) leverage this feature and constructed a full coverage CFG for the CFQ language, which provides us with a basis of source grammar.

Although the CFG is linguistically reasonable, our transduction grammar must simultaneously account for both English and Japanese languages (Wu, 1997); thus we revise this monolingual CFG to satisfy the necessity for translation. We are aware that arbitrary modification especially pruning can cause inconsistency with the well-implemented CFG, accordingly we adopt an "extensive" strategy to create the transduction rules. Here we present a grammar example within this extended framework.

GRAMMAR

$$\begin{aligned} \text{VP} &\rightarrow \langle \text{V NP},\ \text{NP V} \rangle \\ \text{V} &\rightarrow \langle \text{VT andV},\ \text{VT andV} \rangle \\ \text{andV} &\rightarrow \langle \text{and V},\ \epsilon\ \text{V} \rangle \\ \text{NP} &\rightarrow \langle \text{a film},\ \overset{\bar{\text{e}}\ \text{ga}}{\text{映画}} \rangle \\ \text{V} &\rightarrow \{ \langle \text{edit}\ , \overset{\text{hen}\ \text{shū}\ \text{shi}\ \text{ma}\ \text{su}}{\text{編集します}} \rangle, \\ &\quad\ \langle \text{write}\ , \overset{\text{ka}\ \text{ki}\ \text{ma}\ \text{su}}{\text{書きます}} \rangle \} \\ \text{VT} &\rightarrow \{ \langle \text{edit}\ , \overset{\text{hen}\ \text{shū}\ \text{shi}}{\text{編集し}} \rangle, \\ &\quad\ \langle \text{write}\ , \overset{\text{ka}\ \text{ki}}{\text{書き}} \rangle \} \end{aligned} \tag{6}$$

GENERATED STRING

$$\begin{aligned} &\langle \text{write and edit a film}, \overset{\bar{\text{e}}\ \text{ga}\ \text{wo}\ \text{ka}\ \text{ki}\ \ \text{hen}\ \text{shū}\ \text{shi}\ \text{ma}\ \text{su}}{\text{映画を書き 編集します}} \rangle \\ &\langle \text{edit and write a film}, \overset{\bar{\text{e}}\ \text{ga}\ \text{wo}\ \text{hen}\ \text{shū}\ \text{shi}\ \ \text{ka}\ \text{ki}\ \text{ma}\ \text{su}}{\text{映画を編集し 書きます}} \rangle \end{aligned} \tag{7}$$

Note that except for the relatively intuitive constituent reordering issue, the transduction rules are supposed to handle the more challenging multi-mapping problem at the lexical level. Briefly speaking, we derive new tags (corresponding to multiple lexical mappings) for constituents at the lowest syntactic level where the context accounts for the multi-mapping. (6) and (7) give an example informally.

In the string pair of (7), the Japanese verbal inflection is reasoned from its position in a sequence where correspondences are highlighted with different colors. To make it more intuitive, consider a phrase (out of the corpus) "*run and run*" with repeated verb "run" and its Japanese translation "走り走ります", where the repeated "走り"(which should belong to V if in (6)) refers to a category of verb base, namely *conjunctive* indicating that it could be potentially followed by other verbs[3]; and the inflectional suffix "ます" indicting the end of the sentence. Briefly speaking, in the Japanese grammar, the last verb in a sequence have a different form from the previous ones depending on the formality level.

We notice that in this case, the transduction rule of the lowest syntactic level explaining this inflection is $\mathrm{V} \rightarrow \langle \mathrm{VT\ and V},\ \mathrm{VT\ and V} \rangle$, therefore the VT with *suffix* T is derived from V (V exhibit no inflection regarding ordering in English) from this level and carries this context information down to the terminals. Considering questions with deep parse trees where such context information should potentially be carried through multiple part-of-speech symbols in the top-down process, we let the *suffix* be *inheritable* as demonstrated in (8).

$$\begin{aligned} \mathrm{VP} &\rightarrow \langle \mathrm{VPT\ and VP},\ \mathrm{VPT\ and VP} \rangle \\ \mathrm{VPT} &\rightarrow \langle \mathrm{VT\ NP},\ \mathrm{NP\ VT} \rangle \end{aligned} \tag{8}$$

where suffix T carries the commitment of inflection to be performed at the non-terminal level, was explained by context of VPT and inherited by VT. While such suffix is commonly used in formal grammar, we leverage this mechanism to a large extent to fill the linguistic gap. Our strategy is proved to be simple yet effective in practical grammar construction to handle most of the problems caused by linguistic differences such as inflection as mentioned.

A comparison in statistics between our implemented transduction grammar and English CFG for CFQ is shown as Table 1. We notice that there are 47 more lexical units in Japanese at terminal level compared with English indicating the more complex inflections exhibited by Japanese. We accordingly introduce 48 new terminals (mostly derivations) following our strategy, which also lead to a boost in production rules.

	ORG	**TG**		
	EN	**EN**	**EN-JP**	**JP**
Production Rule	125	253	253	-
Reordered Rule	-	-	62	-
Terminal	75	75	287	122
Non-terminal	61	109	117	112

Table 1: Statistics of our transduction grammar (**TG**) compared with original monolingual English CFG (**ORG**); Note that the terminals in EN-JP transduction grammar are represented as $\langle Term_{EN}, Term_{JP} \rangle$ in pairs, terminals in EN and JP refers to their unique terminals respectively; the numbers of non-terminals are the counts of appeared symbols, and EN-JP represents the union of EN and JP.

4.3 Corpus Generation

With the transduction grammar depicted above, the corpus is generated in a pipeline as Figure 1 shows.

4.3.1 Challenge

Since we largely rewrote the original grammar, *failures* and *ambiguities* frequently occurred when parsing the sentences with our initial transduction grammar. Most of them were due to missing certain associated rules when rewriting one rule, especially for the complex chained rules for verb sequences where we should consider all the potential verb derivation symbols. In addition, our framework requires us to construct the *dictionary* by assigning translation for each potential lexical pairs, which is fairly time consuming.

Lexical gap is another notable problem in rule-based MT, While in the more culture-specific text this gap is rather significant, the questions in CFQ are generally culture-invariant, therefore the problems here are mainly ascribed to the difference in word usage. For example, different from the word "marry" in English, the corresponding word in Japanese "結婚する" is *intransitive* and interpreted closer to "get married", which means it should be applied in a different manner from the other verbs.

4.3.2 Detailed Processes

To address the challenges as mentioned above, in addition to the fundamental transduction framework, we further include several detailed processes in practical dataset generation.

[3]Formally, the conjunctive (連用形) in Japanese involves 2 forms:中止形and テ形, to keep consistent with the English questions (where temporal ordering is not entailed by coordination), we adopt the former form in our grammar since it indicates weaker temporal ordering than the latter (Saegusa, 2006).

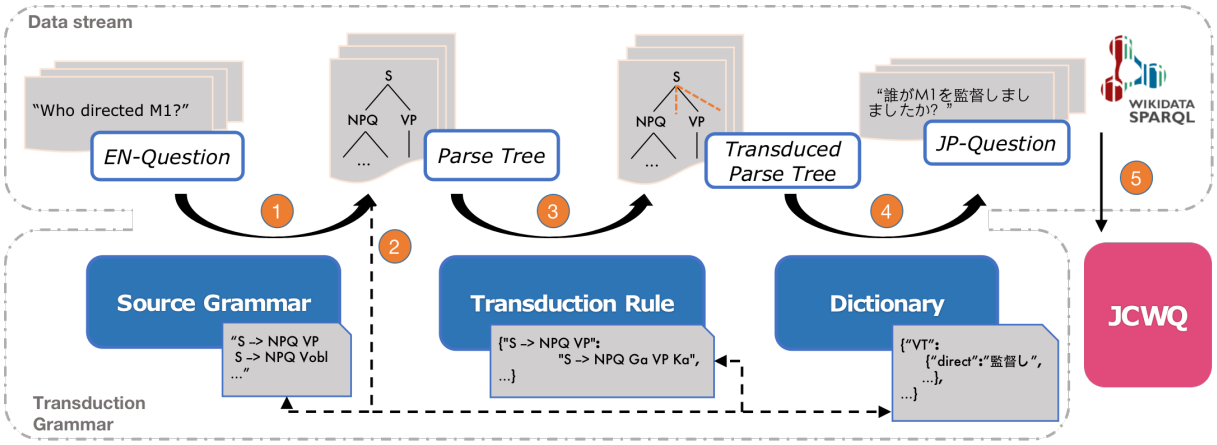

Figure 1: The pipeline of corpus generation. The circled numbers refer to 1. text parsing; 2. build dictionary and revise source grammar and corresponding transduction rules based on parse trees; 3. constituent replacement; 4. lexical translation; 5. post-processing & Wikidata grounding.

Notice that the first two steps were repeated multiple times. We depend on this iterative *parsing&adjusting* process to improve our grammar which causes unexpected errors or ambiguities; meanwhile, we build the dictionary based on the parsing results, i.e, the immediate word-tag pairs occurred in the corpus, therefore the dictionary is specified towards such occurrences (otherwise we have to build a full coverage dictionary for all potential word-tag pairs).

It is worth mentioning that in dictionary construction, we particularly select the lexical translations based on Wikidata[2] to retain consistency with the grounded knowledge base. A typical example for this is the opposite two items *"employer"*[4] and *"employee"*[5] frequently appearing in the questions. While a number of words in Japanese could be selected as a translation, and the entry forming the most natural sentence could vary depending on specific context, we select the only Japanese *label* of the entities as our translations, i.e., "雇用主(*employer*)"[4] and "被雇用者(*employee*)"[5]. Furthermore, the dictionary is strictly injective at the lemma level thus obtaining deterministic translation and retaining the overall atomic distribution. Our strategy for dictionary construction enables us to measure compound divergence in future work.

At the final stage, we found that additional processing is required for disambiguation and better fluency. Specifically, we adopt post-processing primarily aiming at a few semantical ambiguities hard to address with grammar adjustment and issues caused by lexical gaps.

For the ambiguities *systematically* caused by specific ambiguous syntactic patterns, we prescribe *preferred patterns* involving a set of production rules for the automatic selection of translation candidates, i.e., a candidate resulted from a parse tree possessing specific *preferred* structures will be selected in higher priority. This processing handles most of the translation ambiguities. For the lexical gap problem, we adopt a pattern match-substitution strategy to ensure that the context is natural with the "special" terms such as "結婚する" as mentioned in 4.3.1.

Following CFQ and MCWQ, we ground the translated questions in Wikidata through their coupled SPARQL queries; the shared SPARQL queries as in MCWQ enable comparative study with Google translated branches in both cross-lingual and monolingual domains.

5 Evaluation

Our evaluation of the question translations comprises reference-based assessment and manual assessment. The evaluation of the translations focuses not only translation quality as in general MT system, but also how appropriately the translation trade off between the fluency as natural language and failthfulness in the principle of compositionality. For comparison, we further assess the results obtained with Google Cloud Translate as done in MCWQ.

[4] Q3053337
[5] Q703534

5.1 Metric

The most commonly used metric for the reference-based assessment automatically completed by the algorithm is BLEU (Papineni et al., 2002). BLEU measures the translation quality faithfully based on the word-level comparison between predictions and references. SacreBLEU is a variant of BLEU that further addresses the tokenization and normalization scheme difference applied to the references, thus ensuring the BLEU scores are "comparable" (Post, 2018). In this work, we adopt the sacreBLEU package for reference-based assessment.

Although BLEU has been shown to be highly correlated with human assessment results, it explains only the lexical similarity, i.e, semantical consistency which potentially influences the expected answers is not well measured with this method. In addition to the extent to which the translations are satisfying in structural and compositionality consistency, we also measure the general Meaning Preservation (MP) and Fluency (F) ranging from 1 to 5 manually.

We are aware that manual assessment is prone to be subjective, thus the standard was explicitly stated before the assessment. Specifically, we assess the meaning preservation from 2 aspects: whether the lexical units are precisely translated and whether the expected answer domains remained unchanged; intuitively, imprecise lexical translations generally result in an overall score of 3 to 4, while a changed answer domain tends to be penalized more resulting in 1 to 2. For fluency, we assess the naturalness and grammatical correctness of the translations disregarding concrete meaning; unnatural expressions usually result in scores of 3 to 4, while questions with significant grammar mistakes tend to be assigned 1 to 2. Generally, we regard translations with scores ≥ 3 as "acceptable".

5.2 Result

Reference-based Assessment Following the setup of MCWQ (Cui et al., 2022), we sampled 155 questions from the intersection of test sets of 3 MCD splits in CFQ, namely *test-intersection*. We manually translated these questions as reference and call this set *text-intersection-gold*. Note that the gold standard translations are the *expected translations* which we believe are appropriate for parallel compositional corpus.

We calculate the BLEU scores of the 2 versions of translations against the *test-intersection-gold* with sacreBLEU. As shown in Table 2, our method reached 97.1 on BLEU, indicating a nearly perfect translation as we expected. While the rule-based system under full control could ideally reach a full score, the loss here is mainly caused by the samples suffering from problems such as lack of context (or entity) information. In addition, we observe the Google Cloud Translate obtained quite poor performance with 45.1 on BLEU score, which is significantly lower than the other branches in MCWQ (87.4, 76.6 and 82.8 for Hebrew, Kannada and Chinese, respectively; Cui et al., 2022). The main reason for this gap is the different manners in which we set the gold standards. We stress that the reference represents our expectation in a faithful yet possibly rigid way as in original English questions, and not necessarily the most natural expressions. We give detailed analysis in the next section.

Manual Assessment We manually assessed the translation of 42 questions from different structural complexity in terms of meaning preservation and fluency. The results are shown in Table 2. We notice that our translations have significantly better meaning preservation than Google translation, which is exhibited by both the average scores (1.1 higher in avgMP) and the "acceptable rate" (28.6% higher in $\mathrm{P}(\mathrm{MP} \geq 3)$). However, the 2 methods obtained close averaged fluency scores, indicating that both suffer from unnaturalness in translations, partially because of the unnaturalness of original English questions. Furthermore, from the translation samples, we observed that our model produces few translations with significant grammar errors and semantical distortion which influence the expected answer domain, while Google translated results have a noteworthy proportion (28.6%) of unacceptable translations possessing distorted meaning and inconsistent expected answers and a smaller proportion (16.7%) are even grammatically incorrect. We put further analysis with examples in the next section.

6 Results and Analysis

6.1 Generated Corpus

We follow the pipeline in Figure 1 to generate our corpus. The set of the English questions involves **105461** total unique items. [6] The parsing process

[6]Note that we refer to the unique questions with entity **placeholders** here. MCWQ (Cui et al., 2022) reported 124187

Method	Reference	Manual				
	BLEU	avgMP	P(MP $\geq$ 3)	avgF	P(F $\geq$ 3)	P(MP, P $\geq$ 3)
Ours	97.1	4.8	100.0%	4.0	100.0%	100.0%
GoogleTranslate	45.1	3.7	71.4%	4.1	83.3%	71.4%

Table 2: Assessment scores for the Japanese translations. **MP** refers to Meaning Preservation and **F** refers to Fluency. The prefix **avg** indicates averaged scores.

with rewritten source grammar suggested **37280** ambiguities, i.e., our source grammar generated multiple parse trees for these sentences. However, we found most of such ambiguities at the parsing stage are considered in transduction rule construction and "harmless" that they result in single strings as output. Therefore we left these ambiguities to the post-translation stage and found **613** of them are still ambiguous resulting in multiple outputs. The mentioned *preferred pattern* mechanism addressed most of these ambiguities involving regular patterns. A typical example is the sentences beginning with *"Was a film director..."* which the parser occasionally decomposes as *"Was a film director..."* depending on the context (followed by a phrase of verbal past particles for example); a concrete example is shown as below.

Input:*"Was a film director influenced by [Christopher Nolan]"*

Output (Before PP):
{"ある映画監督は*[Christopher Nolan]*に影響され ました か",
"ある映画は*[Christopher Nolan]*に影響された監督でしたか"}

Expected (After PP):
"ある映画監督は*[Christopher Nolan]*に影響されま したか"

in this case, for the post-processing (PP) step, we simply prescribed the correct rule "commonNoun -> P commonNounHead" corresponding to the phrase pair "⟨film director, 映画監督⟩" as a preferred pattern for automatic selection, which take the first candidate as the final output.

However, 322 (0.31%) sentences remaining ambiguous after processing are randomly assigned the final translation from their candidates. We analyzed these ambiguities and here we give some representative examples:

1. *"Who was influenced by a composer influenced by M3 and influenced by M4 and M5 and influenced by M1"*
2. *"Were M3 and M4 written by a composer influenced by M1 and influenced by M2"*

The questions above are even not intuitive for a human to parse. We notice that the 2 sentences are *amphibologies*, indicating that they are grammatically ambiguous. While we can reason that sentence 1 is more fluent with *"influenced by M4 and M5"* rather than *"influenced by a composer... and M5"* as a phrase, and *"M3 and M4"* in sentence 2 are not likely to refer to human, the grammar is unable to conduct further parsing like this. For this small proportion of ambiguities, such reasoning relies on semantical information involving the entities represented as placeholders beyond the scope of grammar.

6.2 Translation Quality Assessment

The sample-based assessment scores indicate the translation quality. Whereas it is generally a comprehensive concept, we argue that for the compositional corpus the quality primarily refers to the consistency in atom, compound, and the overall semantic levels, which is largely manifested by the BLEU and the Meaning Preservation scores in our assessment.

6.2.1 Error Analysis for BLEU scores

As discussed in 5.2, our fully controllable model can translate the sentences as expectations ideally to the largest extent. However, even except for the amphibologies (no amphibological sentences were sampled for assessment) potentially increasing error, our model is still not optimal mainly due to lack of context information. For example, adjective indicating nationalities such as *"American"* is naturally adapted to *"アメリカ人(American person)"* when modifying a person in Japanese; then for a sample (note that entities are bracketed):

Input:*"Was [Kate Bush] British"*
Output:*"[Kate Bush]*はイギリスのでしたか"
Expected:*"[Kate Bush]*はイギリス人でしたか"

unique questions with entity **names**.

Consider the bracketed entity [Kate Bush] which is invisible during translation, and also the fact that the sentence still holds if it is alternated with non-human entities. We notice that without the contribution of the entity semantics, the grammar is unable to specify "人(*person*)" in this case, and results in a less natural expression. We observed a few samples similar to this which caused the error.

Another notable observation is the poor performance of Google Translate on BLEU scores. Except for the "rigid" feature of our references as mentioned in 5.2, another reason is the *multiple grammatical systems* applied by Google Translate. The Japanese language possesses complex honorifics resulting in multiple grammatical systems which are generally distributed in different scenes (Obana, 1991). While we uniformly adopt one system (丁寧体(*polite*)), at least 2 systems (普通体(*plain*), 丁寧体(*polite*)) and further grammatical adaptation (regarding topics) were observed as far as for the samples. Since the Google Translate model infers based on a large training corpus that could cover a wide range of scenarios, the corresponding honorifics (grammatical systems) can also be regarded as its inferences given the text. However, we argue that such scenario-adaptation inference introduces noise to our parallel data and also conflicts with the principle of compositional consistency.

6.2.2 Error Analysis for Manual Assessment

In manual assessment, we disregard our expectations and assess the translations in a general way. As shown in Figure 2, the loss in meaning preservation of our model is largely due to the same reasons as for BLEU scores, i.e., a few lexical translations are not precise enough because of the lack of context, especially entity information. However, as stated in 5.2, our translations cause few semantical distortions potentially changing expected answers, while for Google Translate we observed samples translated as:

Input:"*What did [human] found*"

Output (Google):"*[human]* は何を見つけましたか"

Expected (&Ours):"*[human]*が創設したのは何でしたか"

Disregarding the sentence patterns, the output of Google Translate distorted the meaning as "*What did [human] find*", translated back to English.

Input:"*Was a prequel of [Batman: Arkham*

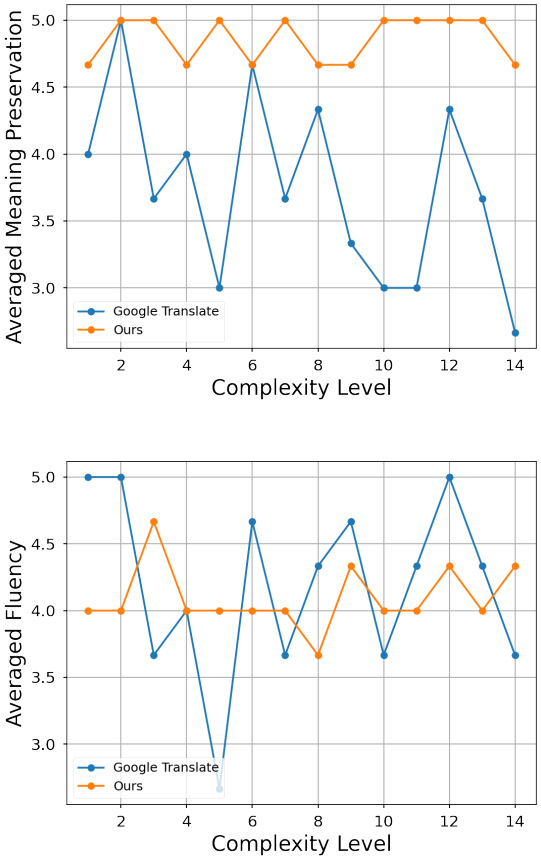

Figure 2: Manual assessment scores vary against increasing complexity level with a bin size of 3, i.e., the scores are averaged over every 3 original complexity levels.

Knight] 's prequel..."

Output (Google):*"[Batman: Arkham Knight]* の前日譚..."

Expected (&Ours):*"[Batman: Arkham Knight]* の前日譚の前日譚..."

The example above shows how the 2 models deal with a compositional phrase occurring in the corpus. Google Translate exhibits reasoning ability which understood that *"a prequel of a prequel"* actually indicates *"a prequel"* thus translating it as *"前日譚(prequel)"*, whereas an expected compositionally faithful translation should be *"前日譚の前日譚(a prequel of a prequel)"*. The 2 representative examples show how Google Translate as a neural model fails in accommodating compositionality even for the well-formed translations: the *infinite* compositional expression potentially reaches the "fringe area" of the trained neural model distribution, i.e., it overly concerns the possibility that the sentence occurs instead of keeping faithful regarding the atoms and their compositions.

To make an intuitive comparison between our model and Google Translate, we divide the 42 complexity levels (for each level we sampled 1 sentence) into 14 coarser levels and see the variation of the scores of 2 models against the increasing complexity. As shown in Figure 2, our model exhibits uniformly good meaning preservation ability while Google Translate suffers from semantical distortion for certain cases and especially for those of high complexity. For the variation of fluency, the steady performance of our model indicates that the loss is primarily *systematic* and due to compromise for compositional consistency and parallel principle, while Google Translate generates uncontrollable results with incorrect grammar (and thus illogical) occasionally.

7 Conclusion

Compositional generalization has been broadly researched on monolingual English benchmarks, and Cui et al. (2022) first bring this topic to a multilingual category through NMT. In this work, we propose to utilize a rule-based methodology for cross-lingual extension of the synthetic corpus and create the Japanese branch as an instance. We further conduct sample-based evaluation to investigate the methods for creating such parallel corpora, from which evidences suggest that our method better accommodates this task in terms of parallel principle. Our work will provide a more reliable benchmark for evaluating cross-lingual compositional generalization, and inspire future studies to create parallel corpora in controllable yet efficient manners.

The controllably generated corpus covering deterministic patterns provides us with access to the atomic and compound distribution thus enabling us to measure the distribution divergence between training and test sets in future work as in CFQ (Keysers et al., 2019). In addition, experiments with state-of-the-art models will also be conducted on our corpus to give a new benchmark as well as compared with the one generated by Google Translate for further analysis.

However, we are aware that further investigation is required in the collection of multi-lingual corpora which serves the multi-lingual NLP development. As Hershcovich et al. (2022) stated, the cultural-awareness should be considered in multilingual NLP research. Although the principle of parallel and compositional consistency impels us to partially sacrifice the naturalness of the corpus, this trade-off is nevertheless worth rethinking in future work.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684.

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.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl' 05)*, pages 263–270.

David Chiang. 2006. An introduction to synchronous grammars.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.

Noam Chomsky. 1965. Aspects of the theory of syntax.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*.

Ruixiang Cui, Rahul Aralikatte, Heather Lent, and Daniel Hershcovich. 2022. Compositional generalization in multilingual semantic parsing over wikidata. *Transactions of the Association for Computational Linguistics*, 10:937–955.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2019. Cross-lingual machine reading comprehension. 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 1586–1595.

Emily Goodwin, Siva Reddy, Timothy O' Donnell, and Dzmitry Bahdanau. 2022. Compositional generalization in dependency parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6482–6493.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, et al. 2022. Challenges and strategies in cross-cultural nlp. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013.

Irina Higgins, Nicolas Sonnerat, Loic Matthey, Arka Pal, Christopher P Burgess, Matko Bošnjak, Murray Shanahan, Matthew Botvinick, Demis Hassabis, and Alexander Lerchner. 2018. Scan: Learning hierarchical compositional visual concepts. In *International Conference on Learning Representations*.

Tsung-Yuan Hsu, Chi-Liang Liu, and Hung-Yi Lee. 2019. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. 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 5933–5940.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. 2019. 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 *2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 9087–9105. Association for Computational Linguistics (ACL).

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Daniel Marcu and Daniel Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 133–139.

Truong-Phat Nguyen. 2021. Urbans: Universal rule-based machine translation nlp toolkit. https://github.com/pyurbans/urbans.

Yasuko Obana. 1991. A comparison of honorifics in japanese and english languages. *Japanese Studies*, 11(3):52–61.

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.

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, Belgium, Brussels. Association for Computational Linguistics.

Reiko Saegusa. 2006. Hanashi kotoba ni okeru tekē (Te form in spoken Japanese language). *Hitotsubashi University Center for Student Exchange Journal*, 9:15–26.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Philip Williams, Rico Sennrich, Matt Post, and Philipp Koehn. 2016. Syntax-based statistical machine translation. *Synthesis Lectures on Human Language Technologies*, 9(4):1–208.

Shuly Wintner. 2016. Translationese: Between human and machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 18–19.

Dekai Wu. 1996. A polynomial-time algorithm for statistical machine translation. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 152–158.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational linguistics*, 23(3):377–403.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *NAACL-HLT*.