Compositional Generalization in Machine Translation for Low-resource Languages

Yutong Liu

Master of Science
School of Informatics
University of Edinburgh
2022

Abstract

This project is about compositional generalization ability of low-resource languages NMT. Compositional generalization means “learn infinite from finite”. For a low-resource language NMT which doesn’t have much training resources, compositional generalization is very important. For now, most of related researches are concentrated on compositional generalization ability of high-resource language NMT.

This project will construct a compositional generalization test suite and test it on the English-Tamil NMT, English-Gujarati NMT and English-German NMT. First of all, this project constructs three kinds of data sets: natural, semi-natural and synthetic. Using these as initial datasets, a compositional generalization test suite is constructed. Finally, NMTs are tested using this test suite.

Experiments in this project validate the lack of compositional generalization ability of NMT and the tendency of NMT to have global compositionality rather than local compositionality as assumed in many studies. It also proves that NMT may have compositional generalization ability but cannot use it correctly. NMT systems trained on more parallel data seem to have better local compositionality.

Research Ethics Approval

This project does not involve any human subjects, and all datasets, tools, and models used are published publicly. This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Yutong Liu*)

Acknowledgements

I would like to thank the University of Edinburgh and my supervisor for their guidance and help. It was their encouragement that helped me complete the dissertation.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

NMT (neural machine translation) is a collective name for a series of machine translation algorithms using neural networks. With its excellent performance, NMT algorithms are widely used in the field of machine translation. At present, NMT has been widely commercialized. It helps people to understand articles in other languages and promotes cross-cultural communication.

However, neural network algorithms are data-driven, which means its performance depends on the quantity of data. More data means better results. For NMT, a large number of parallel corpora are the key to ensure the quality of the model. For high resource language pairs, such as English-Chinese, English-Dutch, NMT can obtain excellent results with the support of a tremendous amount of data; For low resource language pairs, such as English-Tamil, due to the lack of corpus, the translation results of NMT are significantly worse than those of high resource language pairs. The deficiency of data is a difficult problem that low resource language NMT cannot overcome.

In order to improve performance of NMT models for low resource language pairs, it is necessary to understand the behavior of NMT models. Compositional generalization is the ability to use learned language combinations to generate new language combinations. Current research shows that NMT model is lack of compositional generalization ability. The sentences that NMT can learn is finite, while the possible combinations in natural languages are infinite. NMT models with good compositional generalization ability are more likely to obtain better translation results when the data scale is limited. It can be seen that the compositional generalization ability is very important for low resource language NMT. Therefore, studying and evaluating the compositional generalization

ability of a NMT model is of great significance for improving the NMT performance for low resource language pairs. It can help to improve the model structure and build NMT models which are more suitable for low resource languages in the future.

NMTs' compositional generalization ability is related to the interpretability of NMT model. There are few related studies, and most of them are concentrated in high resource languages. The compositional generalization test for English-Tamil NMT (English-Gujarati NMT, English-German NMT) proposed in this project is innovative and can help understand the compositional generalization behavior of low resource language NMT, which is of great significance.

1.2 Objectives

The project will analyze the compositional generalization ability of the English-Tamil NMT model, and compared with the results of English-Gujarati NMT model and English-German NMT model. The project will use the current method for testing high resource language NMT's compositional generalization ability and put it into the English-Tamil setting, with the goal of building an English-Tamil NMT model compositional generalization test suite to deeply understand the compositional generalization behavior of the model.

The objectives of the project mainly include:

1. Collect papers and researches on the interpretability of NMT model and the compositional generalization of NMT model, and use relevant methods to analyze the English-Tamil NMT model.

2. Use relevant methods, construct a test suite to test the English-Tamil NMT model's compositional generalization ability.

3. Use the test suite to test an English-Tamil NMT model, an English-Gujarati NMT model and an English-German NMT model, and obtaining the translation results. Analyze the compositional generalization behavior of those models.

1.3 Results and Outcomes

The project will obtain an English-Tamil compositional generalization test suite, which can be used to evaluate and understand the compositional generalization ability of

NMT model. At the same time, the project will test an English-Tamil NMT model, analyze its compositional generalization behavior and the possible causes of translation results through the test suite.The results are compared with English-Gujarati NMT and English-German NMT. A preliminary evaluation is made on the compositional generalization ability of the low resource language NMT.

The results of this project can also be used to improve the model structure or to study the compositional generalization ability of other language pairs in the future. And finally help to improve hte performance of low resource language NMT model.

1.4 Structure of Dissertation

This paper will be divided into the following five parts:

1. The first chapter is an introduction, which mainly introduces the motivation of the project, the objectives of the project, and the results of the project. This chapter shows some general description of the project, reveals the innovation and importance of the project, and mentions the final goal of the project.

2. The second chapter is the background, which mainly introduces the previous studies and works in the relevant fields of the project. This chapter presents the research and technical background of the project, so that readers can become familiar with the project and get a better understanding of works done by this project.

3. The third chapter is the design and implementation of the project. This chapter introduces each step of the project's implementation and the overall structure of the project in detail. It also elaborates on the construction of the test suite and the testing and evaluation process of the three NMT models. This chapter will introduce the technology in the project combined with the actual work. This chapter will give the reader a comprehensive overview of what the project has done.

4. The fourth chapter is the result and evaluation. This chapter will introduce the compositional generalization test suite constructed by the project and the results of the testing of English-Tamil NMT model,English-Gujarati NMT model, English-German NMT model, and analyze the compositional generalization behavior of NMT models based on the results. This chapter presents the final results

of the project to the readers and proposes understandings of the compositional generalization ability of NMT based on the results.

5. The fifth chapter is a summary, which will summarize the contributions and achievements of the project, and look forward to the following influence of the project to future research.

Chapter 2

Background

2.1 Compositional Generalization

Compositional generalization is the ability to learn unlimited combinations from limited resource[8]. In 1988, Fodor et al.[11] proposed systematical compositionality, pointing out that the ability to understand a complex combination is related to the ability to understand other complex combinations. Hupkes et al.[14] put forward five kinds of compositionality experiments from various definitions and studies, which are as follows: Systematicity, Productivity, Substitutivity, Localism, and Overgeneralisation.

Systematicity experiments focues on testing whether the model can make new combinations that were not included during training. Productivity experiments focues on testing whether the model can handle longer sentences than it was trained on. Substitutivity experiment focuses on testing how the model judges two words as synonyms. Localism experiment focuses on testing whether the compositionality shown by the model is local or global. Overgeneralisation experiment focues on observing the phenomenon of a model to overgeneralize.

2.1.1 Testing on Compositional Generalization

Lake et al.[17]proposed the SCAN dataset based on the above definition to test the compositional generalization ability of neural networks. The dataset consists of a sequence of instructions that the neural network is tasked with translating them into a series of actions. When the neural network translates new commands synthesized from basic command elements, Lake et al.[17] observed the lack of compositional generalization, and pointed out that even if the neural network has learned a method for

systematic compositionality, it still cannot use this ability properly. At the same time, Lake et al.[17] also conducted a small-scale experiment on NMT to test its performance on a new word ”dax”. The result was that the model did not translate the new word well.

After that, Loula et al.[21] used SCAN on RNN (Recurrent Neural Network). This experiment focused on getting the neural network to recombine existing elements rather than learn new elements. This experiment demonstrates that even though RNNs have the ability for compositionality, they are not systematic. Lake et al.[18] also used SCAN for meta sequence-to-sequence learning experiments, pointing out that meta-seq2seq has excellent ability on compositional generalization.

Keysers et al.[15] proposed DBCA (Distribution-Based Compositionality Assessment) and CFQ (Compositional Freebase Questions) datasets. DBCA is a method for evaluating whether a dataset is suitable for measuring compositional generalization, which states that a dataset used for testing compositionality should have a similar distribution of elements with training set and different distribution of combinations with training set. CFQ is a natural language understanding dataset for evaluating compositional generalization and is a more difficult task compared to SCAN.

Akyürek et al.[4] proposed R&R, a learned data augmentation scheme. R&R resamples the excellent and rare compositional examples generated in the original models' prediction, and adds them to the model training data in later training to improve the compositional generalization ability of the model.

Kim et al.[16] proposed COGS, a semantic parsing dataset. The model which is trained on training set of this dataset must have compositional generalization ability to achieve good performance on the generalization set of COGS. Kim et al. conducted experiments and found that the neural network performed poorly on the COGS task. They also found that structural generalization is difficult to lexical generalization.

According to recent researches, neural network have a lot of shortcomings in compositional generalization. This means that it is difficult for them to learn new language combinations from known datasets. Even with the ability of compositionality, it lacks systemicity.

2.1.2 Testing on Compositional generalization for NMT

Currently, researches on NMT compositional generalization are mainly concentrated in high-resource languages. The effect of different training scales on compositional

generalization is mentioned by Dankers et al.[10].

The current research proves that the compositional generalization ability of the NMT model is not excellent. Compositional generalization sometimes happens, but it is not done correctly.

2.1.2.1 English-Chinese

Li et al.[20] proposed a compositional generalization test set of English-Chinese, CoGnition. The data set guarantees to use simple elements and rich combinations to form sentences, and has a sufficient scale to train NMT. Li et al. Built two test sets, one is the common test set, and the other is the compositional generalization test set. The compositional generalization test set uses original words to come up with new combinations. They designed the combination template, embeds the existing elements into the template to construct new combinations, and finally constructs new sentences. In order to construct parallel corpora, Li et al. used the post editing method. They obtain the translation of new sentences with machine translators and then let humans edit them.

Li et al. use transformer to build the tested model. When the generalization test set is used, BLEU decreased significantly, which proves that the compositional generalization ability of transformer has obvious shortcomings. Li et al. continued to analyze other factors. They found that the less a combination appears in the training set, the greater the probability of model making errors. The error probability of zero shot is close to 30%. The longer the combination, the more errors would be made. In the training set, the more frequently words co-occur, the higher the probability that their new combination will be correctly understood by the model. Different combinations, such as NP and VP, have different error probabilities. The frequency of the original words has little effect on the prediction of the model.

2.1.2.2 English-Dutch

Dankers et al.[10] believe that when studying whether a model has compositional generalization ability, it is not appropriate to just consider synthetic data, such as CoGnition[20] and SCAN[17]. Dankers et al. refer to the five experiments proposed by Hupkes et al.[14] and design three experiments: systematicity, substitutivity and overgeneralisation. At the same time, Dankers et al. also considered the difference between local and global generalization, pointing out that most previous studies assumed that compositional generalization occurred locally, which is not the case in natural

language. Dankers et al. used natural datasets of different scales to train the model, and constructed natural, semi-natural, and synthetic datasets as test sets.

In the systematicity experiment, Dankers et al. consider two different sentence combinations: S $\rightarrow$ NP VP and S $\rightarrow$ S CONJ S. Under the first combination, the nouns in the NP of the semi-natural and synthetic datasets are replaced to form new data; the nouns in the VP of the synthetic dataset are replaced to form new data. Under the second combination, different synthetic data are used as the first clause, and the second clause is from the three test sets. Dankers et al. proposed that systematicity guarantees a consistent understanding of the same combination in different contexts, thus using consistency as the evaluation criterion for this experiment. Consistency scores for all settings are low, which means that the model tends to have a global compositional generalization.

In the substitutivity experiment, Dankers et al. used different names for the same thing in British and American English as synonyms, and collected 20 synonym pairs. This experiment also used consistency as the evaluation criterion. It obtained results which is similar to the previous experiment.

In the overgeneralisation experiment, Dankers et al. collected some idioms. Dankers et al. determine whether the translation is literal by detecting keywords in the sentence. It was observed that with the increase of training epochs, the model gradually learned the literal translation of idioms, and then began to learn its true meaning.

Dankers et al. proposed that experiments demonstrate that models with more training data has a better ability of compositionality. Sometimes models will be more compositional and other times it will not, models cannot adjust the compositional generalization at different scales very well.

2.2 English-Tamil NMT

In recent years, most of the NMT model research has focused on high-resource language pairs, such as English-Chinese[7], English-French[25], etc. There are relatively few NMT studies on English-Tamil, and its performance is far less than NMT for high-resource language pairs.

Choudhary et al.[9] proposed to use a combination of word embeddings and BPE (Byte-Pair Encoding)[12] to improve the performance of NMT when encountering untrained words. The model uses Bi-LSTM (Bi-directional Long Short-Term Memory) as encoder and LSTM (Long Short-Term Memory) as decoder, combined with word

embedding, BPE, attention mechanism and other methods to achieves better results than Google Translator.

Ramesh et al.[23] proposed the Samantar dataset, a parallel corpus of Indic Languages. Ramesh et al. then used the samanantar dataset to train on a range of models and examine the performance of the models. The results show that the model trained with Samantar has better translation results.

Although these studies have greatly improved the level of existing English-Tamil translations. They are still incomparable with NMTs in English-Chinese, English-French and other high-resource languages.

Chapter 3

Design and Implementation

3.1 Overall Structure

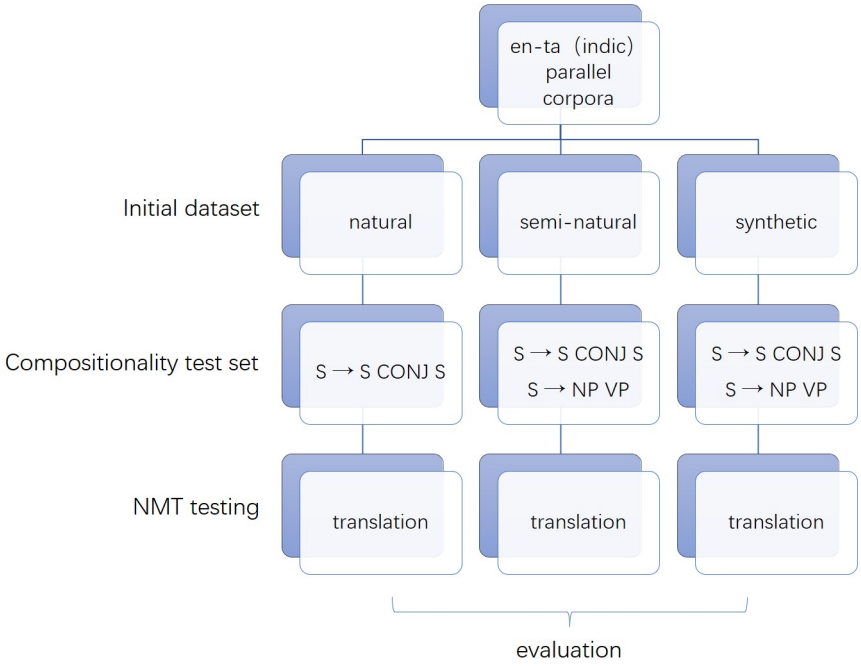

Figure 3.1: Project structure

This project adopts the design of Dankers et al.[10] and applies their method to compositional generalization testing of NMT for indic language. This project mainly refers to the first experiment of Dankers et al.[10], the systematicity experiment. This project is divided into four parts:

The first part is the initial dataset construction. This part will construct three datasets: natural, semi-natural, and synthetic. In this project, these datasets are all in English,

sampled from English-Tamil parallel corpora.

The second part is the compositional generalization test set construction. Using the three datasets obtained in the previous step, datasets for testing compositional generalization was constructed. This step will obtain two different types of data: S $\rightarrow$ NP VP and S $\rightarrow$ S CONJ S.

The third part is the NMT compositional generalization testing. Using English-Tamil, English-Gujarati and English-German NMT as tested models, and using the data obtained in the previous step as test sets. The translation results of the model are collected.

The fourth part is evaluation. This project use consistency[10] as an evaluation criterion. Consistency is calculated from collected results. The obtained consistency scores are then being evaluated and analyzed.

3.2 Experiments' Environment

This project uses python3.7 for programming and uses Pycharm as the IDE platform. The operating system is Windows 10 and Ubuntu 20.

This project uses NLTK, iNLTK[1] and other libraries for assistance. NLTK is an open natural language processing library, and iNLTK is a natural language processing library for indic languages. At the same time, this project uses the code used to synthesize artificial data in the research of Lakretz et al.[19], and, the syntax analysis library, disco-op[24]. Disco-op needs to be operated in a Linux environment.

The English-Tamil[5] and English-Gujarati NMT model[6] under test for this project come from the GoURMET project. They were ran in Docker. English-German NMT model comes from a open source project, hugging face.

3.3 Initial Dataset Construction

From the work of Dankers et al. and Hupkes et al., most compositional generalization experiments consider to use synthetic data rather than natural data, so this project continues this idea by constructing multiple datasets to study different data's impact on NMT's performance on compositionality. The initial dataset contains three datasets: natural, semi-natural, synthetic. Each dataset contains 1500 sentences in English.

3.3.1 English-Tamil Parallel Corpora

This project uses PMIndia[13] as the sampled dataset. PMIndia contains parallel corpora in multiple South Asian languages with English, extracted from the website of the Prime Minister of India, mainly including speeches and news materials. This project is mainly aimed at English-Tamil NMT, so the English-Tamil dataset in PMIndia is selected. The English-Tamil parallel corpus includes a total of 39526 pairs of sentences. This dataset is also used as training set by the NMT used for testing in this project.

3.3.2 Natural Dataset

Table 3.1: Example of natural dataset.

Gopalkrishna Gandhi, the Opposition's candidate, got 244 votes, while 11 votes were declared invalid.
He had served as a senior judge in the Supreme Court of India.

PMIndia has been used as the training set by the tested NMT of this project, so PMIndia is not considered when making the natural dataset. As PMIndia contains a large amount of news data, the project decides to use other Indian news data to produce natural datasets. Indian Express is a large local news website in India which contains multiple Indian languages. The data from this website is chosen as the source of natural dataset for this project. The corpus comes from a public project [2]. The natural dataset is composed of individual sentences with length of 10 to 18 words selected from the corpus, as Table 3.1.

3.3.3 Synthetic Dataset

According to Keysers et al.[15], synthetic datasets need to have a similar vocabulary distribution compared with the training set. In this project, the vocabulary used in the synthetic dataset comes from the high-frequency vocabularies in PMIndia.

The sentences in PMIndia are segmented and the part-of-speech are identified. Classifying according to different part-of-speech, vocabularies of different part is listed in descending order of frequency. Nouns, adverbs and verbs from PMIndia were used

to construct synthetic dataset in this project. Finally, 20 nouns about people, 10 nouns about location, 18 verbs, and 9 adverbs were selected for synthetic data.

Table 3.2: Template using in synthetic dataset. The first column is the name given to the template by Lakretz et al.

nounpp	The N P the N V the N.
simple	The N V the N.
simple_adv	The N adv V the N.
objrel_that	The N that the N V V the N.
subjrel_that	The N that V the N V the N.

Based on the research of Dankers et al.[10], this project used the method of Lakretz et al.[19] to create a synthetic dataset, which was originally used to study the syntactic processing mechanism of LSTM. There are five templates used in this project, as shown in Table 3.2. Each template generated 300 sentences, and a total of 1500 sentences were generated, as Table 3.3.

Table 3.3: Example of synthetic dataset.

nounpp	The brother behind the airport changes the person.
simple	The brother expects the consumer.
simple_adv	The farmer certainly appreciates the president.
objrel_that	The brother that the chief greets congratulates the farmer.
subjrel_that	The brother that appreciates the worker understands the citizen.

3.3.4 Semi-Natural Dataset

For semi-natural datasets, this project constructed them by selecting high-frequency segments in PMIndia. This project counted and extracted language fragments using the disco-op library[24], which itself was developed to analyze discontinuous constituents in natural language.

First, use the parser function of disco-op library to parse the sentences in PMIndia to obtain the sentence structure and form a treebank. After obtaining the tree bank, use the fragment function of disco-op library to count the number of different fragments in the treebank, and output the statistics to a file. This step needs to turn off the discontinuous analysis function of disco-op library. From the obtained statistical data, high-frequency NP (noun phrase) and VP (verb phrase) fragments are selected for subsequent construction of semi-natural datasets. The fragments selected in this project are shown in Table 3.4.

Table 3.4: Fragments selected in this project

NP	(NP (NP) (SBAR (S (VP (TO to) (VP (VB) (NP (NP) (PP (IN) (NP))))))))
NP	(NP (NP (NN)) (PP (IN of) (NP (NP) (PP (IN) (NP (NP) (PP (IN) (NP)))))))
NP	(NP (NP) (PP (IN) (NP (NP (NP) (PP (IN of) (NP))) (CC and) (NP))))
NP	(NP (NP (DT the) (NNS)) (PP (IN) (NP (NP) (PP (IN of) (NP)))))
NP	(NP (NP) (SBAR (S (WHNP (WP)) (VP (VBP) (VP)))))
VP	(VP (MD will) (ADVP) (VP (VB) (NP (NP) (PP))))
VP	(VP (VBN been) (VP (VBN) (PP-CLR (IN) (NP (NP) (PP)))))
VP	(VP (VBG) (NP (NP (DT) (JJ) (NN)) (PP (IN) (NP))))

Use the treesearch function of disco-op library to search the sentence part corresponding to the fragments and output them to a file to obtain sentence parts that have the most common syntactic structure in PMIndia.

After obtaining sentence parts, embed the sentence parts into the template to obtain semi-natural data. The template used in this project and the example of semi-natural data is shown in Table 3.5.

Table 3.5: Example of semi-natural dataset.

The N VP.	The woman will also carry the same spirit of dedication to do something ever-lasting for our nation.
The N speaks about NP.	The leader speaks about production of hydrocarbons beyond the present term of PSC.
Some information about NP is spoken by the N.	Some information about the people on the occasion of Navreh is spoken by the student.
Did the N talks about NP ?	Did the president talks about those who have lost their near and dear ones in this natural calamity?

3.4 Compositional generalization test set

When this project constructed the compositional generalization test set, referring to the research of Dankers et al.[10], two new combinations at the sentence level were considered: S → NP VP and S → S CONJ S. When constructing the first combination, synthetic data and semi-natural data are mainly used; when building the second combination, all three types of data are used.

3.4.1 S → NP VP

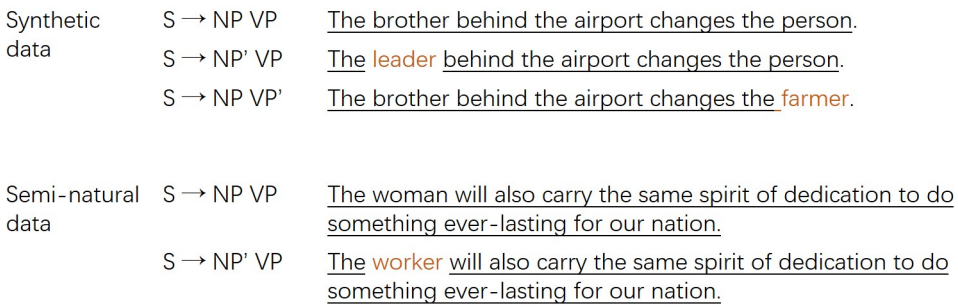

Figure 3.2: Construction of S → NP VP datasets.

When constructing a new combination S → NP VP, this project considers two ways of constitution. One is S → NP' VP, which replaces a noun in NP in the original sentence by a new noun to generate a new sentence, denoted as $S_{NP'}$. The second is S → NP VP', which replaces a noun in VP in the original sentences by a new noun to

generate a new sentence, denoted as $S_{VP'}$. For S → NP' VP, use synthetic data and semi-natural data to generate; for S → NP VP', only use synthetic data to generate.

When generating S → NP' VP, this project randomly replaces the subject in the initial sentence with one of the 20 selected nouns about people mentioned in section 3.2.3. During the generation process, when encountering the template "Some information about NP is spoken by the N" in semi-natural data, replacing the last noun in this template according to Dankers et al.[10].

When generating S → NP VP', this project randomly replaces the object in the synthetic data with one of the 20 selected nouns about people.

See Figure 3.2 for examples of generated datasets. A total of three datasets are generated: semi-natural data $S_{NP'}$, synthetic data $S_{NP'}$, and a synthetic data $S_{VP'}$, 1500 sentences per dataset.

3.4.2 S → S CONJ S

Synthetic data	
S → S1 CONJ S2	The brother that the chief greets congratulates the farmer, and the brother behind the airport changes the person.
S → S1' CONJ S2	The employee that the chief greets congratulates the farmer, and the brother behind the airport changes the person.
S → S3 CONJ S2	The Indian expects the consumer, and the brother behind the airport changes the person.
Semi-natural data	
S → S1 CONJ S2	The brother that the employee appreciates expects the president, and the sister will definitely help me in this task.
S → S1' CONJ S2	he farmer that the employee appreciates expects the president, and the sister will definitely help me in this task.
S → S3 CONJ S2	The employee greets the man, and the sister will definitely help me in this task.
Natural data	
S → S1 CONJ S2	The brother that the teacher changes promotes the soldier, and he had served as a senior judge in the Supreme Court of India.
S → S1' CONJ S2	The employee that the teacher changes promotes the soldier, and he had served as a senior judge in the Supreme Court of India.
S → S3 CONJ S2	The person encourages the leader, and he had served as a senior judge in the Supreme Court of India.

Figure 3.3: Construction of S → S CONJ S datasets.

When constructing a new combination S → S CONJ S, this project considers three ways of constitution. One is S → S1 CONJ S2, connect the original sentence (S2)

and a synthetic sentence (S1) with 'and' to generate a new sentence, denoted as S_{S1}. The second is S $\rightarrow$ S1' CONJ S2, connecting the original sentence and a synthetic sentence different from S1 using 'and' to generate a new sentence, denoted as $S_{S1'}$. There is only one noun that is different between S1' and S1. The third is S $\rightarrow$ S3 CONJ S2, connecting the original sentence with a synthetic sentence that is completely different from S1 to generate a new sentence, denoted as S_{S3}. Each setting uses natural, semi-natural, synthetic three kind of data.

S1, S1', and S3 also come from synthetic data produced by this project. In order to ensure that S1 and S3 are completely different sentences, the order of synthetic datasets is adjusted for building new sentences. The three datasets, S1 and S3 and synthetic data, contain the same content, but in a completely different order. S1' is obtained from S1 using the method for forming S $\rightarrow$ NP' VP in the previous section.

The generated data sample is shown in Figure 3.3. A total of nine datasets were generated: natural data S_{S1}, natural data $S_{S1'}$,natural data S_{S3}, synthetic data S_{S1}, synthetic data $S_{S1'}$, synthetic data S_{S3}, semi-natural data S_{S1}, semi-natural data $S_{S1'}$, semi-natural data S_{S3}. Each dataset contains 1500 sentences.

3.5 NMT Testing

This project uses English-Tamil NMT[5], English-Gujarati NMT[6] and English-German NMT as tested models. English-Tamil NMT and English-Gujarati NMT models are from the GoURMET project[3]. English-German NMT are from hugging face project.

Among them, the parallel corpus used by English-Tamil NMT is about 340K, and the parallel corpus used by English-Gujarati NMT is about 1.1M (including English-Gujarati parallel corpus 42K, and English-Gujarati parallel corpus that translated from Hindi-English parallel corpus 1.1M), according to the integration report of Gourmet project. The training scale of English-Gujarati NMT reaches the "small" setting studied by Dankers et al., while the training scale of English-Tamil NMT is far less than the "small" setting. Since both languages are low-resource languages, data augmentation, back-translation and other methods are used to efficiently use the data when training the model. According to the GoURMET project's report, the best BLEU score of English-Tamil NMT is 11.63, while the best BLEU score of English-Gujarati NMT is 16.4, proving that the translation quality of English-Gujarati NMT is better than the English-Tamil NMT. The English-German NMT model, which is a high-resource

language pair, uses OPUS as the training corpus, and has a much higher training scale than English-Tamil NMT and English-Gujarati NMT. The English-German NMT's best BLEU score on the dataset is 47.3.

English	The brother behind the airport changes the person.
Tamil	விமான நிலையத்தின் பின்புறத்தில் உள்ள அண்ணன் அந்த நபரை மாற்றுகிறார்.
Gujarati	એરપોર્ટ પાછળ ભાઈ વ્યક્તિ બદલે છે.
German	Der Bruder hinter dem Flughafen verändert die Person.

Figure 3.4: Example of translation result

To test the compositionality, the above two models is used to translate the obtained 12 compositional generalization datasets, as well as synthetic and semi-natural data for comparison, into Tamil, German and Gujarati. A total of 21,000 Tamil translation results, 21,000 Gujarati translation results and 21,000 German translation results were obtained (1,500 per dataset). An example of the translation result is shown in Figure 3.4.

3.6 Evaluation

The traditional evaluation method for machine translation is manual evaluation. Since this method is too cumbersome, Papineni et al. introduced an evaluation criterion: BLEU[22]. BLEU is essentially a weighted N-gram matching method. The more N-gram phrases a translation has similar to the standard translation, the better. In some compositional generalization studies, such as Li et al., BLEU is used as an evaluation criterion. BLEU was also used to evaluate the performance of NMT in the previous section.

This project uses the consistency proposed by Dankers et al.[10] as the evaluation criterion. Dankers et al. have observed that if the model has systematicity, it is necessary to ensure that an expression is understood in the same way in different contexts.

In the S→S CONJ S dataset, consistency tests whether S2 is consistent across three different settings. In the S → NP VP dataset, consistency tests when a word changes, whether the rest of the translations in the sentence remain the same.

The formula for calculating consistency is as follows:

$$Consistency = \frac{c_1 + c_2 + \cdots + c_T}{T} \tag{3.1}$$

Among them, T represents the total number of valid translation results in a certain dataset; c_n represents whether the n-th translation result is consistent. When it is S→S CONJ S dataset, if the translation of S2 in this translation is the same as the translation of S2 in S→S1 CONJ S2, $c_n = 1$. If it is different, $c_n = 0$. When it is an S → NP VP dataset, if the translation differs by only one word from the translation of the corresponding initial data, $c_n = 1$, $c_n = 0$ in other cases. The higher the consistency, the more systematic the model is, which means it has a stronger compositional generalization ability.

Compared with the Dutch language studied by Dankers et al., Tamil and Gujarati have more complex conjunction systems, sometimes expressing conjunctions in the form of suffixes. The suffix varies from word to word. Therefore, during the evaluation, the translation results of the S → S CONJ S dataset are uniformly split by punctuation (like “,”, “.”, “;”), and the latter conjunct is identified as S2. If the identification fails, it will not be counted as a valid translation result. For German translation results, use ”und” to separate two sentences. At the same time, German has different definite articles such as ”das” and ”die”. When testing consistency, different articles are uniformly replaced.

Chapter 4

Results and Evaluation

4.1 Dataset Overview

This project firstly constitutes natural, semi-natural, and synthetic, three initial datasets, and then uses these three datasets to constitute compositional generalization test sets. The synthetic data set is constructed by PMIndia's high-frequency vocabulary; the semi-natural data set is constructed by PMIndia's high-frequency syntactic structure; and the natural data set is constructed by other news parallel corpora.

Table 4.1: Dataset Statistics.

Dataset	average length	vocabularies
Synthetic	7.4	98
Semi-natural	18.5	2789
Natural	14.1	5526

Among them, the average sentence length of the semi-natural dataset is the longest, followed by the natural dataset, and the average sentence length of the synthetic data set is the shortest. Natural datasets have the most abundant vocabulary, followed by semi-natural datasets, and synthetic datasets have the least vocabulary, as shown in Table 4.1. It can be seen that the natural dataset is a richer dataset, and the semi-natural data set lacks the richness of natural data although it has a longer sentence length. Synthetic data does not contain very rich information. However, since the artificial dataset does not come from the natural language environment, it has many illogical or rare combinations, see Table 4.2.

According to the above analysis of the dataset, it can conclude that compared with the natural dataset, although there are less vocabularies in artificial dataset, the combination of words in them are very novel. Artificial datasets are very commonly used in the research on compositional generalization of neural networks. The first reason is to better control the distribution of words, and the second reason is to obtain a new distribution of word combinations.

Table 4.2: "Abnormal" sentences in artificial datasets.

Synthetic	The person that ensures the minister appreciates the farmer.
Synthetic	The citizen that the leaders remember promotes the person.
Semi-natural	The police will significantly improve connectivity to Gurugram.
Semi-natural	The sister reflecting a broad convergence of long-term political economic and strategic goals.

4.2 Analysis of Results

This section will discuss the consistency scores obtained from the experiments, and then analyze the compositionality generalization behavior of indic language NMT.

4.2.1 Evaluation for S $\rightarrow$ S CONJ S Setting

In the S $\rightarrow$ S CONJ S experiment, there are two different comparison experiments in total. Figure 4.1 shows the consistency results of S2 when comparing the translation of S_{S1} and $S_{S1'}$, and Figure 4.2 shows the consistency results of S2 when comparing the translation of S_{S1} and S_{S3}.

As can be seen from the two figures, the consistency scores of English-Tamil NMT and English-Gujarati NMT are not high. Compared with the study by Dankers et al.[10], the consistency of both NMTs is lower than that of the English-Dutch model in all aspects, even lower than its minimum training size model. This proves the deficiency of indic language NMT in systematicity, that is, it does not have sufficient compositional generalization ability of and tends to do global compositionality. This may be due

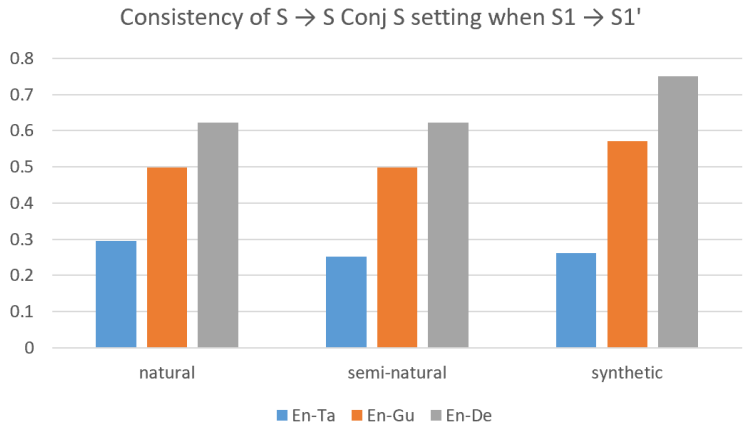

Figure 4.1: Consistency of S_{S1} and $S_{S1'}$

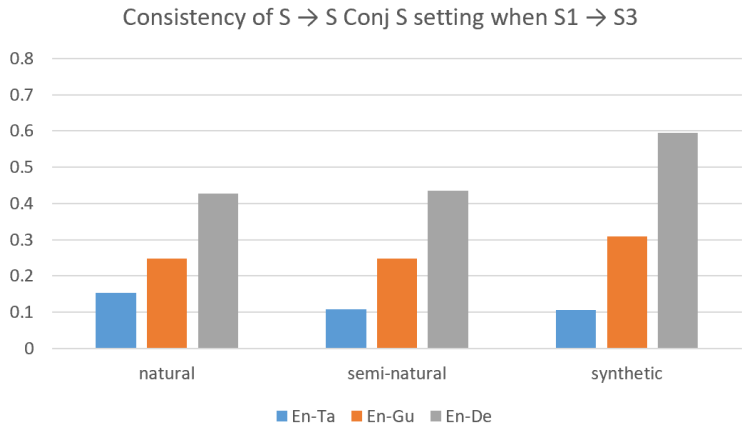

Figure 4.2: Consistency of S_{S1} and S_{S3}

to the lack of effective parallel corpora for indic languages, or the gap between indic languages and European languages. The English-German NMT scores are better, which verifies that the high resource language NMT has better synthetic compositional ability.

The consistency of the S1 $\rightarrow$ S1' experiment is significantly higher than that of the S1 $\rightarrow$ S3 experiment, in both NMTs. A similar phenomenon also exists in the study of Dankers et al. Since the gap between S1' and S1 is significantly smaller than the gap between S1 and S3, this also proves that the model tends to do global compositionality. When the global difference expands, the consistency is also affected. The English-Gujarati NMT has generally higher consistency scores than the English-Tamil NMT. This proves a conclusion similar to Dankers et al. that NMT models with more training data are more prone to local compositionality.

In the English-Gujarati NMT and English-German NMT, it was found that the

consistency scores for the natural and semi-natural data were very similar, while the synthetic data scored significantly higher than the former two. In the English-Tamil NMT, the consistency scores of the natural data were found to be higher than those of the semi-natural and synthetic data, and the consistency scores of the semi-natural and synthetic data are similar. In the study of Dankers et al., natural data and semi-natural data have similar consistency scores, proving that a certain degree of data adjustment does not greatly affect the results of experiment. In this experiment, the results of the English-Gujarati NMT and English-German are in full agreement with the narrative of Dankers et al. Although the result of English-Tamil NMT is not the same with the results of Dankers et al., overall, the differences in the three data for English-Tamil NMT are smaller compared with the other two models's results. This can also prove that a certain degree of control has little effect.

From the data of the three sets of experiments, we can find that the performance of the English-Gujarati NMT and English-German NMT in the experiment are closer to the English-Dutch model of Dankers et al. This may be a consequence of the amount of training data, the parallel corpus used by the English-Gujarati NMT for training reaches the minimal setting of Dankers et al. and English-German NMT have more trainning data compared with English-Gujarati NMT. At the same time, this may cause by language itself. Gujarati, English, German and Dutch all belong to the Indo-European language family, while Tamil belongs to the Dravidian language family. This difference in language classification may also lead to differences in NMT performance.

4.2.2 Evaluation for S → NP VP Setting

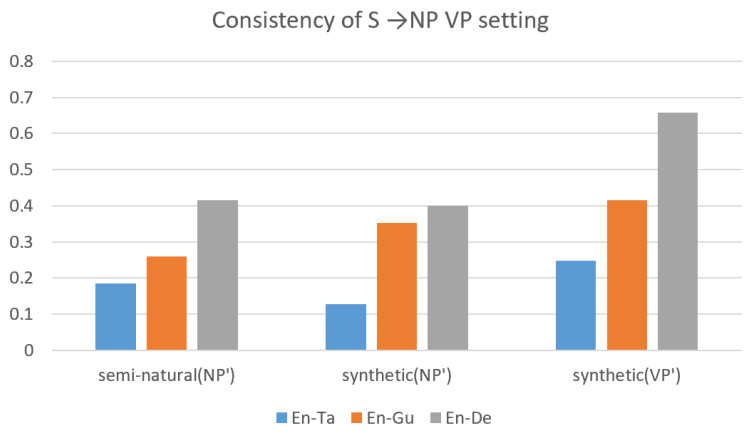

Figure 4.3: Consistency of S → NP VP Setting

In the S→NP VP experiment, there are two different comparison experiments. Figure 4.3 shows the consistency score obtained when comparing the sentence which has a noun changed in NP with the original sentence; and the consistency score obtained when comparing the sentence which has a noun changed in VP with the original sentence.

The experimental results of S→NP VP basically verify some conclusions obtained in the S→S CONJ S experiments. Both English-Tamil NMT and English-Gujarati NMT lack systematicity and tend to do globally compositionality. Compared to the English-German NMT model and the English-Dutch model of Dankers et al.[10], both English-Indic models used in this project scored lower than the minimal training setup. At the same time, the scores of English-Gujarati NMT are higher than those of English-Tamil, which proves that the model trained with more parallel corpora is more inclined to local compositionality. Since natural data was not used in the S→NP VP experiments, it is not clear how much artificially synthesized data affects the experiments. However, the scores of the two English-Indic models are basically consistent with the tendency in S → S CONJ S experiments' results. For English-Tamil NMT, the synthetic data had lower consistency scores than the semi-natural data, but not much different overall. For English-Gujarati, synthetic data's consistency scores are higher than semi-natural data, which is the same with Dankers et al.'s research.

In this experiment, it can be observed that for synthetic data, changing NP has a greater impact on consistency than changing VP. This phenomenon is manifested in both NMT models, but not evident in the study by Dankers et al. Since the changed word in NP is the subject of the sentence which has a great influence on the subsequent meaning of the sentence, the impact is greater than the changed object in VP.

4.3 Analysis of Translation Case

This section will use the cases obtained from the experiments to show and analyze how the compositionality generalization of NMT results in translations. This section mainly takes English-Tamil NMT as an example.

4.3.1 S → S CONJ S Examples

English		It is important to convey to them that they will have support.	
Tamil	S → S1 CONJ S2	அவர்களுக்கு ஆதரவு அளிக்கிறான் என்பதை அவர்களுக்கு உணர்த்துவது முக்கியம்.	It is important to make them realize that you are supporting them.
	S → S1' CONJ S2	அவர்களுக்கு ஆதரவு அளிக்கப்படும் என்றும் அமைச்சர் கூறினார்.	The minister said they would be supported.
	S → S3 CONJ S2	அவர்களுக்கு ஆதரவு கொடுப்பது மிகவும் முக்கியமாகும்.	It is very important to support them.

Figure 4.4: Example of S2's translation when S2 is natural data

Figure 4.5 shows an example produced in the experiment. This example is from the S → S CONJ S experiment, where S2 of this example is the natural data. The first line of the Figure is the actual English of the sentence S2. Column 3 shows the results of translating the sentence S2 into Tamil under three different settings. The fourth column is the result of back-translation from Tamil to English.

According to the comparison, it is obvious that the same sentence is translated into different expressions due to the difference of the first conjunct. Among them, under the setting of S → S1 CONJ S2, the translation of S2 is basically correct. Under the S → S3 CONJ S2 setting, the translation of S2 loses some meaning. In the setting of S → S1' CONJ S2, there is some redundant meaning of the translation of S2. In fact, the extra "The minister" probably comes from S1': "The minister that the student approves thanks the minister."

This example demonstrates that English-Tamil NMT models tend to have global compositionality. The absence of translation in S → S3 CONJ S2, and the mix of translation in S → S1' CONJ S2 both demonstrate that NMT has the compositional generalization ability, but it has not been able to use this ability to obtain correct translations.

Another result getting from this example is that, in the S → S CONJ S experiment, the English-Tamil NMT model translates very poorly on synthetic data. See Figure 4.6, which is the translation of S1, S1', S3 in the same example. Almost all of them have problems such as disordered words and wrong word meanings. This also explains why the English-Tamil NMT scored lower on the synthetic data than the English-Gujarati NMT, the English-German NMT and the English-Dutch NMT by Dankers et al.[10]

English	Tamil	Tamil back-translation to English
The man that the student approves thanks the minister	மாணவன் அமைச்சருக்கு நன்றி சொல்ல ஒப்புதல் அளிக்கிறான்	The student agrees to thank the Minister
The minister that the student approves thanks the minister	மாணவி அமைச்சருக்கு நன்றி தெரிவிக்கும் வகையில் ஒப்புதல் அளிக்கிறார் என்றும்	And that the student approves in gratitude to the Minister
The soldier promotes the employee	சிப்பாய் ஊழியரை ஊக்குவிப்பதோடு	Encouraging the soldier staff

Figure 4.5: Example of S1,S1',S3's translation

English		The consumer behind the building greets the chief.	
Tamil	S → S1 CONJ S2	கட்டிடத்தின் பின்னால் உள்ள நுகர்வோரையும் எதிர்பார்க்கிறது.	Expecting consumers behind the building as well.
	S → S1' CONJ S2	கட்டிடத்தின் பின்னால் உள்ள நுகர்வோரையும் எதிர்பார்க்கிறார்.	He also expects consumers to be behind the building.
	S → S3 CONJ S2	கட்டிடத்தின் பின்னால் உள்ள நுகர்வோர் தலைவருக்கு வாழ்த்துக்களைத் தெரிவிக்கிறார்கள்.	Congratulations to the Consumer Leader at the back of the building.

Figure 4.6: Example of S2's translation when S2 is synthetic data

Figure 4.7 shows an example translation of synthetic data S2 for an S → S CONJ S setting. In this example, as in Figure 4.6, there are many errors in the translation. In S → S3 CONJ S2 setting, the meaning captured by the translation is more comprehensive. The translations obtained for S → S1 CONJ S2 and S → S1' CONJ S2 are almost identical, the only difference being the suffix of the last word. From the back-translated English, it can be known that the suffix expresses rich meanings.

4.3.2 S → NP VP Examples

Figure 4.8 shows an example from S → NP VP experiment. This example is the case where NP is changed, and it is semi-natural data. The first row is the original data, and the second row is the data with replaced words. As can be seen from the figure, the translation itself seems to express the meaning correctly, but the two sentences in Tamil are very different, which proves that after changing the word, the expression of the translation has changed significantly. This is similar to the case of S → S CONJ S. Among them, the first sentence has an extra "And", but the word does not appear in the

English	Tamil	Tamil back-translation to English
The worker will also carry the same spirit of dedication to do something ever-lasting for our nation.	மேலும் தொழிலாளி நமது நாட்டுக்காக நித்தியமான ஒன்றை செய்ய அர்ப்பணிப்பு உணர்வுடன் செயல்படுவார்.	And the worker will act with a sense of commitment to do something lasting for our country.
The woman will also carry the same spirit of dedication to do something ever-lasting for our nation.	மாணவி அமைச்சருக்கு நன்றி தெரிவிக்கும் வகையில் ஒப்புதல் அளிக்கிறார் என்றும்	That woman will act with a sense of commitment to do something for our country forever.

Figure 4.7: Example of S → NP VP experiment when using semi-natural data

original English sentence.

English	Tamil	Tamil back-translation to English
The brother near the house knows the chief.	வீட்டுக்கு அருகில் இருக்கும் சகோதரனுக்கு தலைவன் தெரியும்.	The leader knows the brother who is near the house.
The brother near the house knows the soldier.	வீட்டுக்கு அருகில் இருக்கும் சகோதரனுக்கு சிப்பாய் தெரியும்.	The brother near the house knew the soldier.
The citizen near the house knows the chief.	வீட்டுக்கு அருகில் இருக்கும் குடிமகன் தலைவனுக்குத் தெரியும்	The citizen leader near the house knows.

Figure 4.8: Example of S → NP VP experiment when using synthetic data

Figure 4.9 shows an example of an S → NP VP experiment. This example is synthetic data. The first row is the original data, the second row changes the VP, and the third row changes the NP. The first two sentences are basically correct. Although there is a big difference between the English back-translations, the meaning of them are the same, and the Tamil translation is only one word difference. In the case of changing the NP, the meaning can be barely seen from the back-translated English, and the words is in disorder.

When looking at the results of the S→NP VP experiment, it is found that the translation results on artificial data are significantly better than translation results on artificial data in the S→S CONJ S experiment. In the S → S CONJ S experiment, NMT is more likely to confuse meaning and word order when translating since there are two conjuncts.

These examples all confirm the conclusion above that English-Tamil NMT tends to have global compositional generalization. When NMT using its compositional

generalization ability, it fails to apply properly.

Chapter 5

Conclusion and Future Works

5.1 Conclusion

This project constructs a test suite for testing the compositional generalization ability of English-Tamil NMT. Compositional generalization refers to the ability to understand unlimited language combinations by learning limited language combinations. For low-resource language pair NMT with limited data, the compositional generalization ability is very important. Previous studies have mainly focused on compositional generalization of NMT for high resource languages, but not for English-Tamil NMT.

This project reference the systematicity experiments of Dankers et al.[10]. Natural, semi-natural, synthetic datasets are first constructed. These datasets are used to build compositional generalization test sets. This project considers two sentence-level combinations: S $\rightarrow$ S CONJ S and S $\rightarrow$ NP VP, the first combination has three settings, while the second has two settings. The test sets obtained in this project is used to test both English-Tamil NMT, English-Gujarati NMT, and English-German NMT, using consistency as the evaluation criterion.

For the three different data types, this project proves that a certain amount of data control does not significantly affect the overall experiment. After testing NMTs, this project found that the compositional generalization ability of the two English-Indic NMTs lacked systematicity. Compared with local compositionality, NMT is more inclined to global compositionality. While NMT shows compositional generalization, it lacks the ability to use this ability correctly. English-Gujarati NMT and English-German NMT are stronger than English-Tamil NMT in consistency score, proving that NMT models trained with more parallel data do better on local compositionality.

5.2 Future Works

There is still much room for improvement in this project. In this project, although Tamil and Gujarati are both Indic languages, they belong to different language families. In the future, more low-resource language NMTs can be studied, and the influence of different language families on the compositional generalization ability of NMT can be further studied. Research on more models will also further reveal the impact of the amount of training data on the compositional generalization ability. In this project, the BLEU score of English-Gujarati NMT is higher than that of English-Tamil NMT, in general, more parallel corpora for training means better NMT. Subsequent research can further explore whether it is the model ability or the amount of model training data that affects the compositional generalization of the model.

The compositional generalization test sets can be used to test more NMTs in the future and study their compositional generalization ability. In the future, researchers can be assisted by them in selecting suitable model structures and parameters to bring about better NMT. For low-resource languages, the compositional generalization ability is crucial for NMT. Model structures with excellent compositional generalization ability selected using a compositional generalization test set will benefit NMT for low-resource languages a lot.

Bibliography

[1] https://inltk.readthedocs.io/en/latest/index.html.

[2] https://anuvaad.org/.

[3] https://gourmet-project.eu/.

[4] Ekin Akyürek, Afra Feyza Akyürek, and Jacob Andreas. Learning to recombine and resample data for compositional generalization. *CoRR*, abs/2010.03706, 2020.

[5] Rachel Bawden, Alexandra Birch, Radina Dobreva, Arturo Oncevay Marcos, Antonio Valerio Miceli Barone, and Philip Williams. The university of edinburgh's english-tamil and english-inuktitut submissions to the wmt20 news translation task. In *Proceedings of the 5th Conference on Machine Translation*, pages 92–99. Association for Computational Linguistics (ACL), November 2020. Fifth Conference on Machine Translation, WMT 2020 ; Conference date: 19-11-2020 Through 20-11-2020.

[6] Rachel Bawden, Nikolay Bogoychev, Ulrich Germann, Roman Grundkiewicz, Faheem Kirefu, Antonio Valerio Miceli Barone, and Alexandra Birch. The university of edinburgh's submissions to the WMT19 news translation task. *CoRR*, abs/1907.05854, 2019.

[7] Yong Cheng, Lu Jiang, and Wolfgang Macherey. Robust neural machine translation with doubly adversarial inputs. *CoRR*, abs/1906.02443, 2019.

[8] Noam Chomsky. *Syntactic Structures*. De Gruyter Mouton, 2009.

[9] Himanshu Choudhary, Aditya Kumar Pathak, Rajiv Ratan Saha, and Ponnurangam Kumaraguru. Neural machine translation for English-Tamil. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 770–775, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[10] Verna Dankers, Elia Bruni, and Dieuwke Hupkes. The paradox of the compositionality of natural language: a neural machine translation case study. *CoRR*, abs/2108.05885, 2021.

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

[12] P. Gage. A new algorithm for data compression. *The C Users Journal*, 1994.

[13] Barry Haddow and Faheem Kirefu. Pmindia - A collection of parallel corpora of languages of india. *CoRR*, abs/2001.09907, 2020.

[14] Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. The compositionality of neural networks: integrating symbolism and connectionism. *CoRR*, abs/1908.08351, 2019.

[15] 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. Measuring compositional generalization: A comprehensive method on realistic data. *CoRR*, abs/1912.09713, 2019.

[16] Najoung Kim and Tal Linzen. COGS: A compositional generalization challenge based on semantic interpretation. *CoRR*, abs/2010.05465, 2020.

[17] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2873–2882. PMLR, 10–15 Jul 2018.

[18] Brenden M Lake. Compositional generalization through meta sequence-to-sequence learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[19] Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. The emergence of number and syntax units in LSTM language models. In *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 11–20, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[20] Yafu Li, Yongjing Yin, Yulong Chen, and Yue Zhang. On compositional generalization of neural machine translation. *CoRR*, abs/2105.14802, 2021.

[21] João Loula, Marco Baroni, and Brenden M. Lake. Rearranging the familiar: Testing compositional generalization in recurrent networks. *CoRR*, abs/1807.07545, 2018.

[22] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. Bleu: a method for automatic evaluation of machine translation. 2002.

[23] Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162, 02 2022.

[24] Andreas van Cranenburgh, Remko Scha, and Rens Bod. Data-oriented parsing with discontinuous constituents and function tags. *Journal of Language Modelling*, 4(1):57–111, 2016.

[25] Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566, Vancouver, Canada, July 2017. Association for Computational Linguistics.