Research on the Development of Data Augmentation Techniques in the Field of Machine Translation

Zhipeng Zhang, Aleksey Poguda

***Abstract*—Neural machine translation usually requires a large number of bilingual parallel corpus for training, which is very easy to overfit on the training set of small data. Through a large number of experiments, it has been proved that almost all excellent neural network models are trained on large-scale datasets. High quality bilingual parallel corpus is difficult to obtain, and manual labeling of corpus is usually expensive, and it takes a lot of time. The data augmentation method is an effective technique for scaling data and has achieved significant results in some areas. For example, in the field of computer vision, training data is often augmented with methods such as cropping, flipping, bending or color transformation. Although data augmentation methods have become a basic technique for training neural network models in the field of computer vision, this technology has not been well applied in the field of natural language processing. This article systematically reviews the development of data augmentation techniques in the field of natural language processing in recent years, especially in the subfield of machine translation and conducts research on the mainstream data augmentation methods in the field of machine translation.**

***Keywords*—Natural Language Processing (NLP), Machine Translation (MT), Data Augmentation.**

I. Introduction

With the development of computer technology, machine translation methods have also experienced a long research process. In recent years, the research of neural networks has brought new solutions to machine translation. The application of seq2seq model has made a qualitative leap in the performance of machine translation. The training of neural network machine translation model depends on large-scale bilingual parallel data, which contains sufficient knowledge for machine learning. The training process is the process of data representation and knowledge extraction. How to use data augmentation methods to make model learning easier and knowledge extraction more sufficient is an important research topic.

As one of the core of AI at present, the quantity and quality of data play an almost decisive role in the final performance of a model. The same holds true for machine translation tasks. Since neural machine translation is a supervised learning technology and has super learning ability, the quality and scale of bilingual parallel corpus will be directly related to the final learning effect of the machine translation model. Therefore, before the training of neural machine translation model begins, there are a large number of processing technologies that need to be carried out for data corpus, and a new batch of data can be obtained after processing the original data to support the training of neural machine translation (NMT) models, so as to obtain better learning results. However, building high quality parallel data manually is a costly thing, and it is almost impossible to meet the current demand for data volume of neural machine translation. Therefore, people try to build large-scale parallel data at low cost by automatically building parallel data. At present, the commonly used methods mainly include data mining technology [1] and data augmentation techniques [2]. At present, most people think that data mining technology is a specific technical implementation of data augmentation techniques, so this article includes data mining techniques in data augmentation techniques. Data mining technology mainly uses semantic representation similarity (such as cosine distance of sentence vector) to mine potential parallel data from their respective monolingual corpus. Data augmentation techniques usually uses existing translation models to generate monolingual corpus to obtain synthetic parallel data.

This article will study the application of mainstream data augmentation technologies in machine translation tasks, and summarize and analyze all technologies.

II. Common Scenarios for Data Augmentation Techniques

A. Lack of Samples

In the scenario of few samples, the number of samples that can be collected does not meet the needs of model training, resulting in the model being in the state of under-fitting. Naturally, on the basis of existing data, using data augmentation techniques to expand the sample set is a fast, economical and cost-effective thing. Many studies have also shown that this method can significantly improve the performance of the model [3]-[5].

B. Uneven Sample Distribution in Text Classification Tasks

In addition to some benchmarks, the number of samples of each category in most text classification tasks in real scenarios is uneven. In many cases, the number of categories with the largest number of samples may be two orders of magnitude higher than the number of categories with the smallest number of samples. This will lead to many problems. For example, the model is often in the state of under-fitting for small sample categories. In actual prediction, it will hardly give too high probability for this category [6].

Usually, in the face of such problems, a common way to deal with such a problem is to use data augmentation techniques to expand samples for small sample categories, thereby reducing the imbalance between samples and improving the generalization ability of the model. This

method has also been proven effective many times in practice.

There are many workarounds for sample imbalance. At present, Google's proposed smote method can already solve this problem [7].

C. Semi-supervised Training

From the semi-supervised learning algorithm UDA released by Google, it can be seen that data augmentation techniques can be used on unlabeled samples to construct pairs of samples required for semi-supervised training, so that the model can obtain the gradient required for optimization from unlabeled data [8].

D. Improve the Robustness of the Model

Data augmentation techniques can be divided into two categories without rigor, one is to transform the expression form text while keeping the semantics unchanged, such as the next mentioned back translation, text retelling and etc. The other is to make local adjustments to the original text according to a certain strategy, such as the substitution of synonyms mentioned later, random deletion and etc. Either way, it can be considered to improve the robustness of the model, make the model pay more attention to the semantic information of the text, and is no longer sensitive to the local noise of the text.

Based on this consideration, whether it is a few-sample scenario or a large corpus scenario, the data augmentation techniques help to improve the robustness of the model and improve its generalization ability. On this point, a similar view is expressed in section 7.4 of the book [9].

III. Typical technical solutions

A. Back Translation

Thanks to the remarkable progress in the field of text translation in recent years and the open source of various advanced translation models (including the open interface of translation tools such as Google Translate), data augmentation based on the back translation method has become a general-purpose data augmentation technique with high quality and almost no technical threshold. The basic process of the back translation method is simple, using the translation model to translate the original text of language 1 into the text expression of language 2, based on the expression of language 2 and then translate the text expression of language 3, and finally directly translate the text expression back to language 1 from the language 3, which is the text augmentation by the original text. Of course, many times only one intermediate language can achieve good enhancement results.

Let's use Google Translate for example:

The original text is: Технология улучшения данных в настоящее время является исследовательским направлением машинного перевода.

Translated into Japanese: データ拡張技術は現在、機械翻訳の研究分野です。

Japanese translated into English: Data augmentation techniques are currently a research area of machine translation.

Translate English back to Russian: Методы увеличения данных в настоящее время являются исследовательской областью машинного перевода.

It can be seen that because Google Translate is good enough, the text before and after the augmentation is basically semantically consistent. Therefore, for the augmentation technology of back translation, the quality of the translation model determines the final effect of data augmentation.

If you use a translation model, you can use strategies such as random sample or beam search to achieve exponential data enrichment. If you use translation tools such as Google Translate, you can also achieve N-fold data enrichment by changing intermediate languages.

At present, translation models have weak support for long text input, so in practice, the text is generally split into sentences according to punctuation, and then back-translated separately, and finally assembled into new text.

In the early days, back translation technology was mainly used to improve the performance of neural network translation models[10][11], and monolingual data could be constructed into bilingual data through back translation, thereby helping the model improve performance. Experiments have shown that back translation can help the neural machine translation model bring an average performance improvement of 1.7 BLEU, and help Facebook's team achieve SOTA performance at the time on the WMT'14 English-German test set. The specific implementation process is discussed in detail in the literature [11].

In 2018, the CMU and Google brain teams separated back translation as a specialized data augmentation technique to optimize the performance of question answering models. They trained two neural machine translation models at the same time, English to French and French to English, to achieve back translation, and the specific implementation process is shown in the figure below (fig. 1).

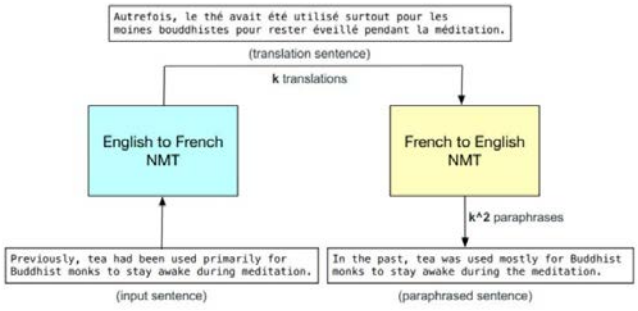

Figure 1. Process Implementation Flowchart

The final experiment proved that back translation technique helped their model achieve at least one percentage point of performance improvement, as shown in the red box in the table below (tab. 1). As we all know, for the question answering system, it is also very good to be able to improve by one percentage point.

Table 1. Method Implementation Effect Comparison List

Single Model	Published[2] EM / F1	LeaderBoard[2] EM / F1
LR Baseline (Rajpurkar et al., 2016)	40.4 / 51.0	40.4 / 51.0
Dynamic Chunk Reader (Yu et al., 2016)	62.5 / 71.0	62.5 / 71.0
Match-LSTM with Ans-Ptr (Wang & Jiang, 2016)	64.7 / 73.7	64.7 / 73.7
Multi-Perspective Matching (Wang et al., 2016)	65.5 / 75.1	70.4 / 78.8
Dynamic Coattention Networks (Xiong et al., 2016)	66.2 / 75.9	66.2 / 75.9
FastQA (Weissenborn et al., 2017)	68.4 / 77.1	68.4 / 77.1
BiDAF (Seo et al., 2016)	68.0 / 77.3	68.0 / 77.3
SEDT (Liu et al., 2017a)	68.1 / 77.5	68.5 / 78.0
RaSoR (Lee et al., 2016)	70.8 / 78.7	69.6 / 77.7
FastQAExt (Weissenborn et al., 2017)	70.8 / 78.9	70.8 / 78.9
ReasoNet (Shen et al., 2017b)	69.1 / 78.9	70.6 / 79.4
Document Reader (Chen et al., 2017)	70.0 / 79.0	70.7 / 79.4
Ruminating Reader (Gong & Bowman, 2017)	70.6 / 79.5	70.6 / 79.5
jNet (Zhang et al., 2017)	70.6 / 79.8	70.6 / 79.8
Conductor-net	N/A	72.6 / 81.4
Interactive AoA Reader (Cui et al., 2017)	N/A	73.6 / 81.9
Reg-RaSoR	N/A	75.8 / 83.3
DCN+	N/A	74.9 / 82.8
AIR-FusionNet	N/A	76.0 / 83.9
R-Net (Wang et al., 2017)	72.3 / 80.7	76.5 /84.3
BiDAF + Self Attention + ELMo	N/A	**77.9/ 85.3**
Reinforced Mnemonic Reader (Hu et al., 2017)	73.2 / 81.8	73.2 / 81.8
Dev set: QANet	**73.6 / 82.7**	N/A
Dev set: QANet + data augmentation ×2	**74.5 / 83.2**	N/A
Dev set: QANet + data augmentation ×3	**75.1 / 83.8**	N/A
Test set: QANet + data augmentation ×3	**76.2 / 84.6**	76.2 / 84.6

In the second half of 2019, the Google team proposed a semi-supervised learning algorithm (UDA) [8] that can be used for NLP tasks, and experimentally proved that data augmentation technique such as back translation can be used for semi-supervised learning, and the results look amazing, they only used 20 samples as label data to achieve near-SOTA performance on the IMDb dataset. At present, back translation is the mainstream data augmentation techniques in the field of machine translation.

B. Random word replacement

The so-called data augmentation method based on random word replacement here is a collective term for a type of text data augmentation method, and its basic method is similar to random cropping and image scaling in image augmentation technique, usually randomly selecting a certain proportion of words in the text, and performing simple operations such as synonym replacement and deletion of these words, unlike models such as back translation, which require the assistance of external pre-trained models.

In 2019, a research team proposed a data augmentation method called EDA (Easy Data Augmentation) [13], which can be considered a collection of such methods. EDA consists of four main operations: synonym replacement, random insert, random swap, and random delete. The detailed description is as follows:

1) Synonym Replacement: Select non-stop words randomly from the sentence. Replace these words with randomly selected synonyms.
2) Random Insert: Randomly find a word in the sentence that does not belong to the stop word set, find its random synonym, and insert the synonym into a random position in the sentence. Repeat n times.
3) Random Swap: Choose two words in the sentence randomly and exchange their positions. Repeat n times.
4) Random Delete: Each word in the sentence is randomly deleted with probability p.

For this method, the biggest doubt is whether the category label of the text can remain unchanged after the EDA operation. After all, this is a random operation on the text. The researchers conducted an experimental analysis specifically on this issue. First, they trained a classification model, let's call it Model A, using only the original training set (without data augmentation). Next, data augmentation is performed on the test set using the EDA method. Finally, the original test set and the expanded corpus are input into model A, and the output of the model at the last linear layer is compared. They found that the distance between the original test set and the expanded corpus is very small in high-dimensional space. The comparison of the results of the two after dimensionality reduction by the t-SNE algorithm is shown in the figure below (fig. 2).

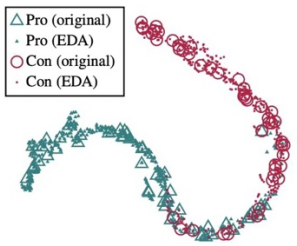

Figure 2. Comparison chart of experimental results

It can be seen from the above analysis that after the EDA transformation, the original data set expands and absorbs a lot of noise on the original basis, expands the amount of data, and maintains the original label, thus effectively expanding the information capacity of the original sample set.

In order to compare more fully, the convolutional neural network (CNN) and recurrent neural network (RNN) were used as classification models, and the average performance of the five tasks was shown in the following table (tab. 2).

Table 2. Result comparison list

	Training Set Size			
Model	500	2,000	5,000	full set
RNN	75.3	83.7	86.1	87.4
+EDA	79.1	84.4	87.3	88.3
CNN	78.6	85.6	87.7	88.3
+EDA	80.7	86.4	88.3	88.8
Average	76.9	84.6	86.9	87.8
+EDA	79.9	85.4	87.8	**88.6**

From the results of the table above, we can draw at least two conclusions:

1) EDA technology can effectively mention the generalization ability of the model, reduce the generalization error, and even under the complete data set. EDA technology can bring an average improvement of 0.8 percentage points.
2) The smaller the dataset, the more significant the model improvement brought by EDA technology. When the sample size is only 500, EDA technology can bring an average improvement of three percentage points. Therefore, it is suitable for scenarios with few samples. It is worth noting that with the help of EDA technology, when the data volume size is only 50% of the original data set, the model performance has exceeded the performance of 100% of the data without EDA.

In summary, we can know that the use of EDA data augmentation technique to improve model performance is simple and effective, especially in small sample scenarios.

C. Non-core Word Replacement

In the EDA technique above, the words to be replaced are randomly selected, so an intuitive feeling is that if some important words are replaced, the quality of the augmentation text will be greatly reduced. The method described in this section is to avoid this problem as much as possible.

The technology was proposed by Google in the article [8]

on the UDA algorithm. The core point of the whole technology is also relatively simple, replacing a certain proportion of unimportant words in the text with unimportant words in the dictionary, thereby generating new text.

In information retrieval, TF-IDF values are generally used to measure the importance of a word to a piece of text, and the following briefly introduces the definition of TF-IDF:

1) Text frequency (TF) is the number of times a word appears in the text, the statistics are word frequency TF, obviously, a word appears many times in the article, then the word may have a great role, but if the word often appears in other documents, such as "of", "I", then its importance is greatly reduced, the latter is to use IDF to characterize.
2) Inverse Document Frequency (IDF) is an importance adjustment factor that measures whether a word is common or not. If a word is rare, but it appears multiple times in this article, then it probably reflects the characteristics of this article and is exactly the keyword we need.
3) TF-IDF = TF×IDF, this formula can effectively measure the importance of a word to a piece of text. When we know the importance of a word to a text, we use the probability of negative correlation with TF-IDF to sample the words in the text to decide whether to replace, which can effectively avoid the wrong replacement or deletion of some keywords in the text.

The specific implementation method proposed in the UDA article is shown in the following figure (fig. 3).

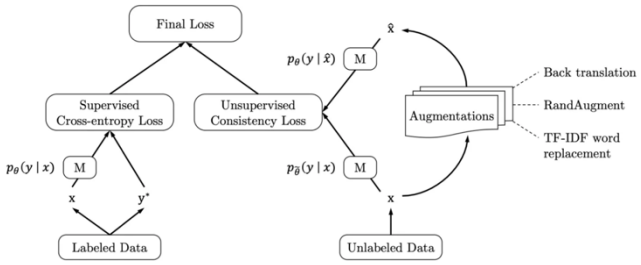

Figure 3. UDA algorithm architecture diagram

The original article proposing this method did not conduct a controlled experiment on this method alone, but worked with the back translation technique to achieve text augmentation. The following table (tab. 3)shows the implementation of the article on six different datasets.

Table 3. Comparative Analysis List of Experimental Results

Fully supervised baseline							
Datasets (# Sup examples)		IMDb (25k)	Yelp-2 (560k)	Yelp-5 (650k)	Amazon-2 (3.6m)	Amazon-5 (3m)	DBpedia (560k)
Pre-BERT SOTA		*4.32*	2.16	29.98	3.32	34.81	0.70
$\text{BERT}_{\text{LARGE}}$		4.51	*1.89*	*29.32*	*2.63*	*34.17*	*0.64*
Semi-supervised setting							
Initialization	**UDA**	IMDb (20)	Yelp-2 (20)	Yelp-5 (2.5k)	Amazon-2 (20)	Amazon-5 (2.5k)	DBpedia (140)
Random	✗	43.27	40.25	50.80	45.39	55.70	41.14
	✓	25.23	8.33	41.35	16.16	44.19	7.24
$\text{BERT}_{\text{BASE}}$	✗	18.40	13.60	41.00	26.75	44.09	2.58
	✓	5.45	2.61	33.80	3.96	38.40	1.33
$\text{BERT}_{\text{LARGE}}$	✗	11.72	10.55	38.90	15.54	42.30	1.68
	✓	4.78	2.50	33.54	3.93	37.80	1.09
$\text{BERT}_{\text{FINETUNE}}$	✗	6.50	2.94	32.39	12.17	37.32	-
	✓	**4.20**	**2.05**	**32.08**	**3.50**	**37.12**	-

In the experiment, four different models were used for control experiments, namely the weight randomized Transformer structure, BERT-base, BERT-large, and the BERT-finetune in the field, and the values in the table are the errors on the test set. It can be seen from the table that after the text augmentation of non-core word substitution and back translation, the model has basically achieved great improvement in each dataset of the experiment.

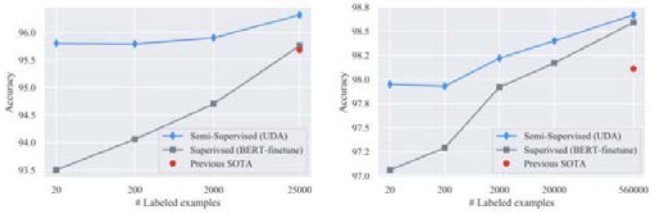

Figure 4. Comparison Chart of Experimental Results of Two Datasets

The figure (fig. 4) above shows the best performance that the model can achieve by using the UDA algorithm and two data augmentation methods under different amounts of labeled data. Regarding the data augmentation technique, an important judgment can be indirectly verified from the figure: whether in a few-sample or large-sample scenario, the use of data augmentation technique can help the model to further improve the performance on the basis of the original sample set.

In this article of UDA, the researchers only used the operation of word substitution, and did not add the other three operations in EDA, such as deletion, swapping positions, etc., which can be used as one of the subsequent research directions.

This technical additional operation is the introduction of TF-IDF to measure the importance of a word to a sentence, which can essentially be considered as the introduction of strong prior knowledge on the basis of EDA, and then replace synonyms according to identified keywords to avoid useless data and erroneous data.

D. Data Augmentation Based on Contextual Information

Data augmentation technique based on contextual information are also simple in principle. First, a trained language model is required, and for the original text that needs to be enhanced, a word or word is randomly removed from the text (depending on whether the language model supports words or words). Next, the rest of the text is input into the language model, and the top k words predicted by the language model are selected to replace the words that have been removed from the original text to form k new texts.

Preferred Networks' contextual data augmentation technology [14] based on bidirectional Language Model was proposed in 2018. The entire architecture is shown in the following figure (fig. 5).

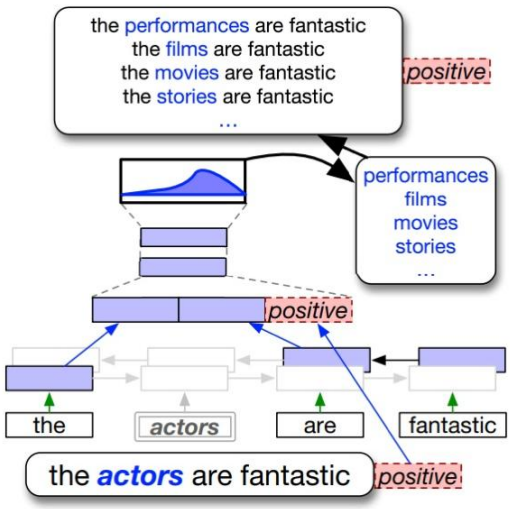

Figure 5. Preferred Networks Corporate' Model Architecture Diagram

In this scheme, compared with the general bidirectional language model, in order to ensure that the label after the text transformation is unchanged. The researchers add in a hidden layer in the language model. The label information of the text is added in, so as to ensure that the generated text has the same label attributes as the original text.

The researchers tested the effectiveness of this method in five classification tasks, and the results are shown in the following table (tab. 4).

Table 4. Comparison list of experimental results

Models	STT5	STT2	Subj	MPQA	RT	TREC	Avg.
CNN	41.3	79.5	92.4	86.1	75.9	90.0	77.53
w/ synonym	40.7	80.0	92.4	86.3	76.0	89.6	77.50
w/ context	41.9	80.9	92.7	86.7	75.9	90.0	78.02
+ label	42.1	80.8	93.0	86.7	76.1	90.5	**78.20**
RNN	40.2	80.3	92.4	86.0	76.7	89.0	77.43
w/ synonym	40.5	80.2	92.8	86.4	76.6	87.9	77.40
w/ context	40.9	79.3	92.8	86.4	77.0	89.3	77.62
+ label	41.1	80.1	92.8	86.4	77.4	89.2	**77.83**

As can be seen from the above table, the method proposed by article can bring about 0.5 improvement compared to the synonym replacement method. However, the question of whether label information should be included is addressed. It can be seen from the experiment that the addition of label information brings a reduction of about 0.2 percentage points of generalization error, which is basically within the fluctuation range of generalization error, so it is doubtful whether there is an obvious effect.

Another remarkable research achievement [15] in this direction comes from the Chinese Academy of Sciences. The overall idea is similar to the above scheme. The main difference is that the bidirectional language model is replaced by BERT, and BERT is also finetuned. The label information of the original text is introduced to ensure that the newly generated sample has the same label attributes as the original sample. The experimental results are shown in the following table (tab. 5).

Table 5. Comparison List of Experimental Results of Chinese Academy of Sciences

Model	SST5	SST2	Subj	MPQA	RT	TREC	Avg.
CNN*	41.3	79.5	92.4	86.1	75.9	90.0	77.53
w/synonym*	40.7	80.0	92.4	86.3	76.0	89.6	77.50
w/context*	41.9	80.9	92.7	86.7	75.9	90.0	78.02
w/context+label*	42.1	80.8	93.0	86.7	76.1	90.5	78.20
w/BERT	41.5	81.9	92.9	87.7	78.2	91.8	79.00
w/C-BERT	**42.3**	**82.1**	**93.4**	**88.2**	**79.0**	**92.6**	**79.60**
RNN*	40.2	80.3	92.4	86.0	76.7	89.0	77.43
w/synonym*	40.5	80.2	92.8	86.4	76.6	87.9	77.40
w/context*	40.9	79.3	92.8	86.4	77.0	89.3	77.62
w/context+label*	41.1	80.1	92.8	86.4	77.4	89.2	77.83
w/BERT	41.3	81.4	93.5	87.3	78.3	89.8	78.60
w/C-BERT	**42.6**	**81.9**	**93.9**	**88.0**	**78.9**	**91.0**	**79.38**

At least two conclusions can be drawn from the experiment:

1) BERT-based contextual data augmentation technique can lead to significant model performance improvements, averaging close to two percentage points, which is still attractive.
2) Bringing the label information of the original text into BERT (W/C-BERT) does bring significant model gains compared to not bringing in (w/BERT).

From the experimental results, the second method has obvious performance improvement compared with the first method, and the BERT model has become the most commonly used pre-training model in the field of NLP, so this method can be easily used by most people.

E. Data Augmentation Based on the Language Generation Model

Data augmentation using language generation model is a large class of methods, and there are currently multiple ways to implement it [16]-[18], and research work before 2019 was generally based on data augmentation techniques derived from RNN architectures for specific tasks. Until 2019, GPT and GPT-2 models were born, and the effect on data generation tasks was extremely amazing.

As a general language generation model that has been pre-trained on massive corpus, GPT will naturally be used to implement data augmentation related work. IBM's research team proposed a text augmentation technique based on GPT architecture in November 2019, which they called LAMBDA (language-model-based data augmentation). LAMBADA first conducted pre-training on a large number of texts to enable the model to capture the structure of the language and thus produce coherent sentences. Then finetune the model on a small number of data sets of different tasks, and use the finetuned model to generate new sentences. Finally, the classifier is trained on the same small dataset and filtered to ensure that the existing small dataset and the newly generated dataset have a similar distribution.

In order to fully verify the performance of LAMBADA technology, researchers conducted two types of experiments.

Experiment 1: LAMBADA technology was applied to three different data sets, and three different model architectures (BERT, LSTM, SVM) were used for the control experiment. The experimental results are shown in the following table (tab. 6).

Table 6. Comparison List of Experimental Results of Three Models

Dataset		BERT	SVM	LSTM
ATIS	Baseline	53.3	35.6	29.0
	LAMBADA	**75.7**	**56.5**	**33.7**
	% improvement	58.5	58.7	16.2
TREC	Baseline	60.3	42.7	17.7
	LAMBADA	**64.3**	**43.9**	**25.8**
	% improvement	6.6	2.8	45.0
WVA	Baseline	67.2	60.2	26.0
	LAMBADA	**68.6**	**62.9**	**32.0**
	% improvement	2.1	4.5	23.0

Baseline refers to the model when only the original data set is used for training. It can be seen from the table that LAMBADA technology can improve performance in all three data sets compared with baseline. Especially for ATIS data sets, the performance of baseline has been improved by more than 50%. The conclusion given in the original article is that ATIS data has obvious uneven distribution, and LAMPADA technology can effectively compensate for the imbalance of the original data set.

Experiment 2: Compare LAMBADA technology with other mainstream data augmentation techniques. The experimental results are shown in the following table (tab. 7).

Table 7. Comparison of Experimental Results of Different Data Augmentation Techniques

Dataset		BERT	SVM	LSTM
ATIS	Baseline	53.3	35.6	29.0
	EDA	62.8	35.7	27.3
	CVAE	60.6	27.6	14.9
	CBERT	51.4	34.8	23.2
	LAMBADA	**75.7***	**56.5***	**33.7***
TREC	Baseline	60.3	42.7	17.7
	EDA	62.6	**44.8***	23.1
	CVAE	61.1	40.9	25.4*
	CBERT	61.4	43.8	24.2
	LAMBADA	**64.3***	43.9*	**25.8** *
WVA	Baseline	67.2	60.2	26.0
	EDA	67.0	60.7	28.2
	CVAE	65.4	54.8	22.9
	CBERT	67.4	60.7	28.4
	LAMBADA	**68.6***	**62.9***	**32.0***

EDA and CBERT have been introduced in detail in the previous. It can be seen from the figure that the advantages of LAMBADA technology are still obvious. If BERT is adopted as the model architecture, it can increase by at least 1.2 percentage points compared with other data augmentation algorithms; In the ATIS data set, it was 13 percentage points higher than the second place. Similarly, in SVM and LSTM, LAMBADA technology is still outstanding, except for individual data and slightly worse performance than EDA.

In a word, at least from the experiment in the article, LAMBADA technology can be regarded as one of the most excellent data augmentation technique at present. LAMBADA technology has a lot to explore in the future, such as combining with the UDA architecture mentioned earlier to realize semi-supervised learning with few samples.

F. New Data Augmentation Techniques

1) Text style transfer

In the field of computer vision, image style transfer has been studied a lot in previous years. For the human eye, although the style of the photo before and after the transformation changes greatly, the person or animal entity on it is still recognizable. In other words, style transfer can also be seen as an image data augmentation.

Along this line, if there are also mature and common language style transfer algorithms in the field of NLP, then naturally they can also be used for text data augmentation. In fact, back translation has a bit of text style transfer, but it belongs to the uncontrollable text transformation. In this regard, articles [19]-[20]have already been published in this regard.

2) Synthetic Translations

Eleftheria Briakou and Marine Carpuat of the University of Maryland, published in ACL2022, “Can Synthetic Translations Improve Bitext Quality?” [21], uses synthetic data to replace imperfectly aligned data in the mined parallel data to obtain high quality parallel corpus.

Recently published corpus through data mining contains a large amount of mistranslated data. Taking WikiMatrix, a recently commonly used multilingual translation corpus, as an example, the article conducted a random sample of Greek to English for manual evaluation, and found that about 12% of the samples had large semantic differences, only certain similarities in topics or structures, while 56% of the samples had fine-grained differences. Only 32% of parallel samples can be perfectly matched.

This article focuses on the analysis of the widespread mismatch of the existing large-scale corpus based on data mining technology, and shows that the synthetic data obtained by the translation model can effectively alleviate this problem.

3) Ciphertext Based Data Augmentation

Data augmentation requires a high quality of the corpus, and the main way to obtain high quality corpus is still human translation. But the training of neural networks often requires a large amount of corpus, and it is unrealistic to achieve this purpose manually. Therefore, it is necessary to find other ways to obtain high quality corpus. The article “CipherDAug: Ciphertext based on Data Augmentation for Neural Machine Translation” [22] presented at ACL 2022 innovatively uses cryptography to obtain high quality corpus.

Author proposes a novel data-augmentation technique for neural machine translation based on ROT-k ciphertexts. ROT-k is a simple letter substitution cipher that replaces a letter in the plaintext with the kth letter after it in the alphabet. Author first generate multiple ROT-k ciphertexts using different values of k for the plaintext which is the source side of the parallel data. Author then leverages this enciphered training data along with the original parallel data via multisource training to improve neural machine translation.

Overall, CipherDAug shows promise as a simple, out-of-the-box approach to data augmentation which improves on and combines easily with existing techniques, and which yields particularly strong results in low-resource settings.

4) Continuous Semantic Augmentation

Supervised learning is limited by the amount of data. Common data augmentation methods cannot generate diverse and faithful samples. This article [23] proposes a new data augmentation method – Continuous Semantic Augmentation (CSANMT), which augments each training instance with an adjacency semantic region that can cover enough literal expression variants in the same sense.

The algorithm addresses two limitations of data augmentation in discrete space: the lack of diversity of augmented training examples in discrete space, and the difficulty of preserving the original semantics of enhanced

text in discrete space.

The algorithm was evaluated on a variety of machine translation tasks, including WMT14 English-German/French, NIST Chinese-English, and multiple IWSLT tasks. Specifically, CSANMT reached the new SOTA in enhanced technology with a score of 30.94 BLEU in the WMT14 English-German task. In addition, the method can achieve comparable performance using a baseline model with only 25% of the training data. This shows that CSANMT has great potential to achieve good results in low-resource situations.

5) Conditional Masked Language Model

The article "Semantically Consistent Data Augmentation for Neural Machine Translation via Conditional Masked Language Model" [24] mainly studies the technology of using word substitution for data augmentation in neural machine translation, which achieves the purpose of data augmentation by replacing words in existing parallel corpus sentence pairs. When using the data augmentation method, the authors observed that if the enhanced data samples retained the correct label information, they could effectively scale the trained data and improve the performance of the model. This property is called semantic consistency.

In neural machine translation systems, training data exists in the form of sentence pairs, including source sentences and target sentences. Semantic consistency requires that both the source and target sentences be fluent and grammatically correct in their respective languages, and that the target sentence should be a high quality translation of the source sentence. Existing word replacement methods are usually swapping, removing, or randomly replacing words in the source and destination sentences. Due to the discrete nature of natural language processing, these transformations do not maintain semantic consistency, and often they may impair the fluency of double sentences or break the correlation between pairs of sentences.

In order to improve the data augmentation method in machine translation training, the semantics of the source and target sentences and the cross-language translation relationship between them can be preserved in the process of enhancement. The Conditional Mask Language Model (CMLM) was introduced, which generates context-sensitive alternate word distributions from which we can choose the best alt for a given word. The CMLM model is a variant of Mask Language Model that incorporates label information when predicting masks.

In order to verify the effectiveness of the proposed method, the authors conducted experimental validation on three smaller datasets: IWSLT2014 German, Spanish, and Hebrew translations to English, and a larger dataset: WMT14 English to German. The experimental results show that the performance effect of CMLM's data augmentation method is significantly better than that of other methods, and the improvement of 1.9 BLEU in WMT English to German has been achieved.

IV. The Effectiveness of Text Augmentation Techniques

A. Regularization

Data augmentation techniques are undoubtedly an effective regularization method, whether it is back translation, EDA, non-core word replacement, or data augmentation based on contextual information, which is essentially a model preference expressed by the designer or imposed a strong prior distribution assumption on the distribution of the model. Among them, the model preference of back translation expression is that the model should have invariance for texts with different forms of expression but the same semantics. The model preference expressed by EDA, non-core word replacement and etc. is that the model should be insensitive to local noise of the text. Therefore, even in the face of few sample scenarios, under this regularization, the model can effectively converge in the hypothetical space and achieve better generalization error.

B. Transfer Learning

Any learning requires effective external information guidance, and the effectiveness of the data augmentation techniques mentioned above can undoubtedly be understood from the perspective of transfer learning. Whether it is back translation, GPT-2 based data augmentation or text style transfer, it can be understood as transferring information or knowledge learned by an externally pre-trained model from elsewhere to the current task, improving the information capacity of the overall data and better guiding the learning of the current model.

C. Improve Model Robustness

EDA and other techniques can not only be viewed from the perspective of semantic noise, but also can be regarded as applying generalized noise (independent of specific tasks) to the input data to achieve functions similar to the dropout layer. This idea has been proven by various studies to improve the robustness of the model to a certain extent.

D. Manifold

The text of the same type of label can be regarded as a kind of manifold in the text space, so effective data augmentation techniques should ensure that the newly generated text is still a point on the manifold.

References

[1] Schwenk H, Chaudhary V, Sun S, et al. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia[J]. arXiv preprint arXiv:1907.05791(2019).

[2] Nguyen X P, Joty S, Wu K, et al. Data diversification: A simple strategy for neural machine translation[J]. Advances in Neural Information Processing Systems, (2020), 33: 10018-10029.

[3] Wei, Jason W., and Kai Zou. "Eda: Easy data augmentation techniques for boosting performance on text classification tasks." arXiv preprint arXiv:1901.11196 (2019).

[4] Anaby-Tavor, Ateret, et al. "Not Enough Data? Deep Learning to the Rescue!." arXiv preprint arXiv:1911.03118 (2019).

[5] Hu, Zhiting, et al. "Learning Data Manipulation for Augmentation and Weighting." Advances in Neural Information Processing Systems. (2019).

[6] Wang, William Yang, and Diyi Yang. "That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets." Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. (2015).

[7] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research16 (2002): 321-357.

[8] Xie, Qizhe, et al. "Unsupervised data augmentation." arXiv preprint arXiv:1904.12848 (2019).

[9] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, (2016).
[10] Sennrich, Rico, Barry Haddow, and Alexandra Birch. “Improving neural machine translation models with monolingual data.” arXiv preprint arXiv:1511.06709 (2015).
[11] Edunov, Sergey, et al. “Understanding back-translation at scale.” arXiv preprint arXiv:1808.09381 (2018).
[12] Yu, Adams Wei, et al. “Qanet: Combining local convolution with global self-attention for reading comprehension.” arXiv preprint arXiv:1804.09541 (2018).
[13] Wei, Jason W., and Kai Zou. “Eda: Easy data augmentation techniques for boosting performance on text classification tasks.” arXiv preprint arXiv:1901.11196 (2019).
[14] Kobayashi, Sosuke. “Contextual augmentation: Data augmentation by words with paradigmatic relations.” arXiv preprint arXiv:1805.06201 (2018).
[15] Wu, Xing, et al. “Conditional BERT contextual augmentation.” International Conference on Computational Science. Springer, Cham, (2019).
[16] Liu, Ting, et al. “Generating and exploiting large-scale pseudo training data for zero pronoun resolution.” arXiv preprint arXiv:1606.01603 (2016).
[17] Hou, Yutai, et al. “Sequence-to-sequence data augmentation for dialogue language understanding.” arXiv preprint arXiv:1807.01554 (2018).
[18] Dong, Li, et al. “Learning to paraphrase for question answering.” arXiv preprint arXiv:1708.06022 (2017).
[19] Hu, Zhiting, et al. “Toward controlled generation of text.” Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, (2017).
[20] Guu, Kelvin, et al. “Generating sentences by editing prototypes.” Transactions of the Association for Computational Linguistics 6 (2018): 437-450.
[21] Eleftheria Briakou, Marine Carpuat “Can Synthetic Translations Improve Bitext Quality?” Published by ACL 2022.
[22] Nishant Kambhatla, Logan Born, Anoop Sarkar. “CipherDAug: Ciphertext based Data Augmentation for Neural Machine Translation” Published by ACL 2022.
[23] Xiangpeng Wei, Heng Yu, Yue Hu, Rongxiang Weng, Weihua Luo, Rong Jin. “Learning to Generalize to More: Continuous Semantic Augmentation for Neural Machine Translation” Published by ACL 2022.
[24] Qiao Cheng, Jin Huang, Yitao Duan. “Semantically Consistent Data Augmentation for Neural Machine Translation via Conditional Masked Language Model”. arXiv preprint arXiv:2209.10875(2022).

ZHIPENG ZHANG was born in China. He received his bachelor's degree from North China University of Science and Technology in 2017. Now he is studying for a master’s degree at Tomsk State University. Email: Zhipeng_Zhang0411@outlook.com

Poguda Aleksey Andreevich is the professor of the Department of Information Support of Innovative Activity of the Faculty of Innovative Technologies of the National Research Tomsk State University. Member of the Council of Young Scientists of TSU, Leading Programmer of the Laboratory of Personal Computers and Multimedia Devices. Email: aapoguda@gmail.com