Dataset construction method of cross-lingual summarization based on filtering and text augmentation

Hangyu Pan [1] , **Yaoyi Xi** [Corresp., 1] , **Ling Wang** [1] , **Yu Nan** [1] , **Zhizhong Su** [1] , **Rong Cao** [1]

[1] State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, China

Corresponding Author: Yaoyi Xi
Email address: WIM_GY@163.com

Existing cross-lingual summarization (CLS) datasets experience inconsistent sample quality and low scale. To address the problems, in this study, we propose a method that jointly supervise quality and scale to build CLS datasets. In terms of quality supervision, the method adopts a multi-strategy filtering algorithm to remove low-quality samples of monolingual summarization (MS) from the perspectives of character and semantics, improving the quality of the MS dataset. In terms of scale supervision, the method adopts a text augmentation algorithm based on the pretrained model to increase the size of CLS datasets with quality assurance. Based on the method, we also build an English-Chinese CLS dataset and evaluate it with a reasonable data quality evaluation framework. The evaluation results show that the dataset is of good quality and large size, which proves that the proposed method can both comprehensively improve the quality and effectively increase the scale, thereby obtaining a high-quality and large-scale CLS dataset at a lower cost.

Dataset construction method of cross-lingual summarization based on filtering and text augmentation

Hangyu Pan[1], Yaoyi Xi[1], Ling Wang[1], Yu Nan[1], Zhizhong Su[1], Rong Cao[1]

[1]State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, Henan, China

Corresponding Author:
Yaoyi Xi
No.62 Science Avenue, Zhengzhou, Henan, 450001, China
Email address: WIM_GY@163.com

Abstract

Existing cross-lingual summarization (CLS) datasets experience inconsistent sample quality and low scale. To address the problems, in this study, we propose a method that jointly supervise quality and scale to build CLS datasets. In terms of quality supervision, the method adopts a multi-strategy filtering algorithm to remove low-quality samples of monolingual summarization (MS) from the perspectives of character and semantics, improving the quality of the MS dataset. In terms of scale supervision, the method adopts a text augmentation algorithm based on the pretrained model to increase the size of CLS datasets with quality assurance. Based on the method, we also build an English-Chinese CLS dataset and evaluate it with a reasonable data quality evaluation framework. The evaluation results show that the dataset is of good quality and large size, which proves that the proposed method can both comprehensively improve the quality and effectively increase the scale, thereby obtaining a high-quality and large-scale CLS dataset at a lower cost.

Introduction

Cross-lingual summarization (CLS) converts *texts*[1] in one language into *summaries* in another language to enable people to quickly and efficiently obtain information from *texts* written in unfamiliar languages. The researches of CLS has evolved from pipeline approaches (Leuski et al., 2003; Siddharthan & McKeown, 2005; Orăsan & Chiorean, 2008; Wan, Li & Xiao, 2010; Wan, 2011; Yao, Wan & Xiao, 2015; Zhang, Zhou & Zong, 2016; Ayana et al., 2018; Wan et al., 2019; Ouyang, Song & McKeown, 2019) to end-to-end approaches (Duan et al., 2019; Zhu et al., 2019; Xu et al., 2020; Cao, Liu & Wan, 2020; Takase & Okazaki, 2020; Ladhak et al., 2020;

[1]we use "text" to refer to a carrier of information in general, alongside the categories such as image and speech, and "*text*" to refer specifically to the input in the sample pair (text-summary) of Automatic Text Summarization, which means that "*summary*" represents the output in the sample pair.

Dou, Kumar & Tsvetkov, 2020; Yin et al., 2020; Zhu et al., 2020; Bai, Gao & Huang, 2021; Bai et al., 2021; Wang et al., 2021), and the end-to-end approach is currently introducing deep learning models, such as Transformer (Vaswani et al., 2017). Extensive work has shown that the quality and scale of annotated data directly affect the performance of deep learning models. Therefore, both the quality and scale of the CLS dataset are extremely important.

Currently, researchers have constructed some CLS datasets through the collection method (Ladhak et al., 2020; Nguyen & Daumé, 2019; Fatima & Strube, 2021) and the transformation method (Ayana et al., 2018; Duan et al., 2019; Zhu et al., 2019). The most representative one is NCLS constructed by Zhu et al. (2019). Datasets obtained by the collection method are of higher quality while the cost is also high, thus they are small in scale. The transformation method builds CLS datasets from datasets of other tasks at a low cost and with a guaranteed scale. However, datasets obtained by the transformation method contain more low-quality samples, which seriously affects the performance of CLS methods. There are two reasons for this phenomenon. First, errors in the source dataset. For example, Zh2EnSum, the subset of NCLS, which is derived from LCSTS (Hu, Chen & Zhu F, 2015), contains many *summaries* that are too abstract because of the characteristics of the microblog, as shown in Table 1. Second, errors in the transformation system, such as translation error. Therefore, building high quality and large-scale datasets at low cost is a serious challenge for CLS research.

To address the problems of existing datasets and their construction methods, in this paper, we propose a dataset construction method of CLS based on filtering and text augmentation that jointly supervises quality and scale. In terms of quality supervision, the method uses the multi-strategy filtering algorithm (MSF) which includes the strategies of irrelevant word statistics, keyword statistics, and semantics measure, to remove low-quality samples of monolingual summarization (MS). In terms of scale supervision, the method uses the text augmentation algorithm based on the pretrained model (TAPT) to increase the size of CLS datasets.

The evaluation results show that MSF can simply and effectively improve the quality of MS datasets, and TAPT can increase scale with assured quality which can be used to both improve the performance of CLS systems and build CLS datasets. The CLS dataset constructed by our method is of extremely high quality and large scale, which indicates that our method can both comprehensively improve the quality and effectively increase the scale, thereby obtaining a high-quality and large-scale CLS dataset at a lower cost.

The main contributions of this paper are as follows.

1. We propose MSF to improve the quality of MS datasets, which removes low-quality MS samples from the perspective of character and semantics. It is the first time to automatically check the degree to which the *summary* reflects the content of its original *text*, and realizes the content comparison between non-parallel texts. The strategy of semantics measure in MSF implements the similarity measure for non-parallel texts, which can be widely applied.

2. We propose TAPT to increase the size of text data with quality assurance. TAPT not only uses the self-attention mechanism, which is good at capturing the internal correlation of data or features, to select the words to be replaced, but also uses MLM, which is an unsupervised pre-

training task of the pretrained model, to realize contextual dynamic synonym replacement, greatly improving the effect of text augmentation. Experimental results shows that fine-tuning MBART (Liu et al., 2020) with TAPT can achieve +19.83 ROUGE-1, +15.4 ROUGE-2, +17.4 ROUGE-L for English-Chinese CLS and +1.49 ROUGE-1, +0.31 ROUGE-2, +4.99 ROUGE-L for Chinese-English CLS compared to the previous best performance (Zhu et al., 2019). TAPT can be used in conjunction with any supervised CLS method to further improve the performance of CLS systems.

3. We propose a general and effective dataset construction method of CLS based on filtering and text augmentation. The method not only guarantees the quality of CLS dataset, but also meets the requirement of its scale. It can be used to build more CLS datasets. In addition, we also applied this method to build a high-quality and large-scale English-Chinese CLS dataset (En2Zh_Sum) with the data size of 2830266, which can be directly used for future research.

Related Works

CLS dataset

Current dataset construction methods of CLS can be summarized as the collection method and the transformation method. The overview of common CLS datasets is shown in Table 2.

The collection method refers to obtaining texts from resource-rich platforms, such as the Internet, and organizing them into CLS datasets. The process is shown in Fig 1. Ladhak et al. (2020) collected multilingual CLS datasets from WikiHow[2]. Nguyen & Daumé (2019) collected multilingual CLS from Global Voices[3]. Fatima & Strube (2021) collected English-German CLS datasets from Spektrum der Wissenschaft[4] and Wikipedia[5].

The transformation method refers to automatically generating CLS datasets from datasets of other tasks through a transformation system. The process is shown in Fig 2. Ayana et al. (2018) built an English-Chinese CLS dataset by translating the *summaries* of Gigaword (Napoles, Gormley & Durme, 2012) and DUC (Over, Dang & Harman, 2007) while Duan et al. (2019) built a Chinese-English CLS dataset by translating the *texts* of Gigaword and DUC. Zhu et al. (2019) built English-Chinese and Chinese-English CLS datasets by translating *summaries* of CNN/Daily Mail (Hermann et al., 2015) and LCSTS (Hu, Chen & Zhu F, 2015), respectively, using a filtering strategy based on ROUGE (Lin, 2004).

Text augmentation

Data augmentation is a method for generating a large amount of data from a small amount of data using semantic invariance as a criterion (Schwartz et al., 2018). Common text augmentation algorithms can be categorized as word-level and text-level. The overview of related researches is shown in Table 3.

In word-level augmentation, Wei & Zou (2019) proposed EDA (Easy Data Augmentation), which includes four operations: synonym replacement, random insertion, random exchange, and

[2]https://www.wikihow.com
[3]https://globalvoices.org
[4]https://www.spektrum.de
[5]https://www.wikipedia.org

random deletion. Kobayashi (2018) proposed a contextual text augmentation that uses a bidirectional language model for contextual dynamic synonym replacement. Wu et al. (2019) replaced the bidirectional language model of Kobayashi (2018) with BERT (Devlin et al., 2018). In text-level augmentation, Yu et al. (2018) used back-translation (BT) (Sennrich, Haddow & Birch, 2016) for text augmentation in reading comprehension tasks. Xie et al. (2019) proposed UDA (Unsupervised Data Augmentation) for unsupervised text augmentation using BT. Some studies used Natural Language Generation (NLG) model for augmentation. Hou et al. (2018) proposed a data augmentation framework based on a sequence-to-sequence (Seq2Seq) model for the text augmentation of dialogue systems. Anaby-Tavor et al. (2019) proposed LAMBDA (Language-model-based Data Augmentation), which used GPT-2 (Radford et al., 2018) to generate new texts for augmentation.

Methods

To address the problems of existing datasets and their construction methods, we propose a dataset construction method of CLS based on filtering and text augmentation. Firstly, the method applies MSF to improve the quality of the MS dataset, whose language is the target language of CLS (*text* in the source language-*summary* in the target language). Secondly, the method translates the *text* of the MS dataset into the source language, and matches the translation with the corresponding *summary* of the original *text* to obtain a CLS dataset. Finally, the method uses TAPT to expand sample pairs of the CLS dataset, so as to obtain a high-quality and large-scale CLS dataset. The method not only guarantees the quality of CLS dataset, but also meets the requirement of its scale. The process is shown in Fig 3.

Multi-strategy filtering

To accurately measure how well the *summary* in MS dataset generalize the *text* content, we propose multi-strategy filtering algorithm. The algorithm uses the strategies of irrelevant word statistics, keyword statistics, and semantics measure successively to remove low-quality MS sample pairs from the perspective of character, combination of character and semantics, and semantics, so as to improve the quality of datasets. The overall process is shown in Fig 4.

Irrelevant word statistics

The words in the *summary* that do not appear in its original *text* (defined as irrelevant words) will affect the learning effect of the CLS model to some extent. Therefore, this strategy calculates the proportion of irrelevant words in the *summary* to all *summary* words to measure how much *text* content the *summary* contains from the perspective of character. If the proportion is too high, it means that there are too many words in the *summary* that do not appear in the original *text*, and the sample should be filtered out.

Specifically, given the *text* of a MS sample $X=\{x_1,...,x_i,...,x_m\}$ and its reference *summary* $Y=\{y_1,...,y_j,...,y_n\}$, m is the length of X, n is the length of Y, $n<m$. x_i and y_j denote the i th word of X and the j th word of Y, respectively. Then the proportion of irrelevant words r_A is:

$$r_A = \frac{|\{y \in Y | y \notin X\}|}{n} \tag{1}$$

where $|\cdot|$ denotes the cardinal number of a set.

Keyword statistics

A good *summary* should contain many keywords of the original *text*. Word embedding can reflect the semantic relationship of words in high-dimensional spaces, and is a good choice for measuring semantic similarity to introduce semantic information (Tang et al., 2019). K-means algorithm (Macqueen, 1966) can cluster similar objects into a same cluster. Therefore, this strategy uses a word clustering method based on the Word2Vec (Mikolov et al., 2013a; Mikolov et al., 2013b) to extract keywords of a *text* from the perspective of semantics, and then calculate the proportion of words in a *summary* belonging to keywords of its corresponding *text* to all words in the *summary* to measure how much key information of the *text* is contained in the *summary* from the perspective of character. If the proportion is too low, it means that the *summary* has too many non-keywords, and the sample should be filtered out.

Specifically, given X and Y, we first encode X with Word2Vec to derive the word representation sequence $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_i, ..., \boldsymbol{x}_m\}$, and cluster all the words with K-means algorithm. Then we calculate the Euclidean distance between the cluster centers and other words, using the cluster centers as the main keywords, and selects the p nearest words to the cluster center as keywords to obtain the keyword set $C = \{c_1, ..., c_p\}$. Then the proportion of *summary* words belonging to keywords of the *text* r_B is:

$$r_B = \frac{|\{y \in C\}|}{n} \tag{2}$$

where $|\cdot|$ denotes the cardinal number of a set.

Semantics measure

A good *summary* should be semantically similar to the original *text*. Contextual word embeddings from the pretrained model, such as BERT (Devlin et al., 2018), have brought a leap forward in semantic representation of texts. However, due to the problem of anisotropy, BERT-based text embedding cannot measure similarity using cosine similarity. BERT-Whitening (Su et al., 2021) solves the problem by transforming the embedding vector into isotropic form by simply whitening (i.e., principal component analysis). Therefore, this strategy takes BERT-Whitening as text embedding, and calculate the cosine similarity between representation vectors of the *text* and its *summary* to measure how much *text* content the *summary* contains from the perspective of semantics. If the cosine similarity is too small, the similarity between the *summary* and the *text* is too low, and the sample should be filtered out.

Specifically, given X and Y, we first obtain the word representation sequences of X and Y by BERT word embedding, $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_i, ..., \boldsymbol{x}_m\}$ and $\boldsymbol{Y} = \{\boldsymbol{y}_1, ..., \boldsymbol{y}_j, ..., \boldsymbol{y}_n\}$ respectively, then obtain their text representation vectors $\boldsymbol{x}'$ and $\boldsymbol{y}'$. Following which, $\boldsymbol{x}'$ and $\boldsymbol{y}'$ are unified and denoted as $\boldsymbol{z}'$. $\{\boldsymbol{z}_k'\}_{k=1}^{2N}$ is whitened and h principal components are retained to obtain $\{\tilde{\boldsymbol{z}}_k'\}_{k=1}^{2N}$.

The process is shown in Table 4 (Su et al., 2021). Finally, $\{\tilde{\boldsymbol{z}}_k'\}_{k=1}^{2N}$ is split into $\left(\tilde{\boldsymbol{x}}_s', \tilde{\boldsymbol{y}}_s'\right)_{s=1}^{N}$, and the cosine similarity r_C between $\boldsymbol{x}'$ and $\boldsymbol{y}'$ is:

$$r_C = \cos\left(\tilde{\boldsymbol{x}}', \tilde{\boldsymbol{y}}'\right) \tag{3}$$

where $\cos(\cdot)$ computes the cosine similarity of two vectors.

Text augmentation based on the pretrained model

Self-attention (Vaswani et al., 2017) can capture inter-word dependencies. MLM, a pre-training task of auto-encoded pre-trained models such as BERT and RoBERTa (Liu et al., 2019), can contextually predict words. Therefore, we propose a text augmentation algorithm based on the pretrained model that uses the self-attention and MLM to dynamically replace synonym words for generating a new *text*.

Specifically, given the *text* of a CLS sample $X^{src} = \{x_1^{src}, \dots, x_i^{src}, \dots, x_m^{src}\}$ and its reference *summary* $Y^{tgt} = \{y_1^{tgt}, \dots, y_j^{tgt}, \dots, y_n^{tgt}\}$, we first use self-attention to select the words to be masked, obtaining $X_{masked}^{src} = \{x_1^{src}, \dots, <mask>, \dots, x_m^{src}\}$. Subsequently, we predict the masked words by using the MLM of pretrained model to obtain the new *text* $X^{src'} = \{x_1^{src}, \dots, x_i^{src'}, \dots, x_m^{src}\}$. Finally, $X^{src'}$ and Y^{tgt} are constructed together as a new CLS sample. The process is shown in Fig 5, where blue text indicates that the predicted result is different from the original *text*, and green text indicates that the predicted result is the same as the original *text*.

Experimental Setup

Dataset

LCSTS (Hu, Chen & Zhu F, 2015) is a Chinese summarization dataset originating from Sina Weibo, containing Part_I, Part_II, and Part_III. The authors scored samples of Part_II and Part_III to judge the relevance of the *summary* to the *text*. The correlation score interval is [1,5], and the higher the score, the more relevant it is. In this study, 2,196,263 samples of Part_I after deduplication and 195 samples of Part_III with a score of 5 after deduplication are used as the original samples for building En2Zh_Sum.

NCLS (Zhu et al., 2019) is the benchmark set of CLS. We use it to validate TAPT. It contains the English-Chinese CLS dataset En2ZhSum and Chinese-English CLS dataset Zh2EnSum. The statistics are shown in Table 5, and the word segmentation algorithm is BPE (Sennrich, Haddow & Birch, 2016). LCSTS is the data source of Zh2EnSum. Due to the large data size, considering the hardware, training effect, training efficiency and other factors, we randomly sample one-sixth of En2ZhSum train set (60,781 samples) and one-half of Zh2EnSum train set (846,857 samples) as the train subsets. And we use TAPT on them to get the augmented train subsets, with the data size reaching 115,589 and 1,424,296, respectively.

Baselines and comparison methods

To validate TAPT, we use it directly for CLS and compare it with some research results. The study of neural CLS is just emerging, and there are not many research results. Some representative research results are as follow.

The following describes the work of Zhu et al. (2019), which is a benchmark for CLS studies and covers pipeline methods and end-to-end methods.

TETran: It translates *texts* in the source language using a transformer-based MT model, and then summarizes the translated *texts* in the target language using the LexRank algorithm (Erkan & Radev, 2004).

TLTran: It summarizes *texts* in the source language using a transformer-based MS model, and then translates *summaries* in the source language to the target language using a transformer-based MT model.

GETran and **GLTran**: It replaces the MT model in TETran and TLTran with Google Translator[6].

NCLS: It trains a Transformer (Vaswani et al., 2017) on NCLS.

NCLS-MT: It trains a Transformer by incorporating MT and CLS under multi-task learning.

NCLS-MS: It trains a Transformer by incorporating MS and CLS under multi-task learning.

The followings are other outstanding CLS studies that have been conducted in recent years.

XNLG-CLS (Xu et al., 2020): It fine-tunes the XNLG model (Chi et al., 2020) on NCLS.

ATS (Zhu et al., 2020): It trains a Transformer on NCLS, then sums the neural network probability distribution of the Transformer and the translation probability distribution of a probabilistic bilingual dictionary as the final *summary* generation distribution.

MLPT (Xu et al., 2020): It pretrains the CLS model using two unsupervised pretraining tasks and three supervised pretraining tasks, then fine-tunes the model by incorporating MS and CLS under multi-task learning.

RL-XSIM (Dou, Kumar & Tsvetkov, 2020): It uses a Transformer to perform multi-task learning for CLS, MT, and MS, and then optimizes the model through bilingual semantic similarity.

MCLAS (Bai, Gao & Huang, 2021): It modifies the output of CLS into sequential connections between MS and CLS.

CSC (Bai et al., 2021): It uses the compression ratio to unify the MT and CLS corpora, and encodes the compression ratio into the semantic representation of *texts*.

The above are the most representative research results of CLS at present. We use them as baselines. The pretrained model BART (Lewis et al., 2020) had achieved state-of-the-art performance on MS at the time, and thus we choose the multilingual pretrained model MBART (Liu et al., 2020) as the basic framework of CLS, and take full advantage of its powerful semantic understanding, cross-lingual alignment and text generation capabilities. Combining the methods in this study, the following three comparison models can be obtained.

MBART-CLS: It uses MBART directly for CLS.

$\mathbf{MBART_{ft}}$-CLS: It fine-tunes MBART on the train subsets of NCLS.

$\mathbf{(MBART{+}TPTA)_{ft}}$-CLS: It fine-tunes MBART on the augmented train subsets of NCLS.

Parameter setup and evaluation metric

Parameter setup

Our dataset construction method belongs to the transformation method. When building En2Zh_Sum, we avoid introducing errors to reference *summaries* that can affect the learning

[6] https://translate.google.com

effect of CLS model by translating *texts* of LCSTS instead of *summaries*, and use Baidu Translate API[7] as the transformation system to ensure translation quality. In MSF, we use jieba[8] library for Chinese word segmentation, while the Word2Vec-based word clustering method is implemented using the Word2Vector of gensim[9] library and K-means algorithm of sklearn[10] library. BERT embedding and whitening are performed using bert-base-uncased[11] of Huggingface-transformers and codes from NLP-Series-sentence-embeddings[12] project. The average word vector of all words in the first and last layers of the BERT word vector is used as text embedding. Li et al. (2020) have proved that this pooling is the optimal choice without any processing. In TAPT, we use BPE (Sennrich, Haddow & Birch, 2016) to tokenize[13] texts and build word dictionary, and put all English texts in lower case. Roberta-base[14] and mbart-large-cc25[15] of Huggingface-transformers[16] are used to implement RoBERTa and MBART.

In the experiments to verify En2Zh_Sum and TAPT, we set the input/output sequence length to 550/100 and 80/60 for English-Chinese and Chinese-English CLS, respectively. The AdamW (Loshchilov & Hutter, 2019) optimizer is used to train in parallel on 2 NVIDIA RTX A6000 GPUs, and we stop fine-tuning after 100,000 iterations. The key parameters of the experiments are shown in Table 6.

To select the most appropriate pretrained model for TAPT, we also test the performance of five classical pretrained models for predicting words, including BERT, ELECTRA (Clark et al., 2020), ERNIE (Sun et al., 2020), RoBERTA and ALBERT (Lan et al., 2020). Specifically, electra-base-discriminator[17], ernie-2.0-base-en[18], and albert-base-v2[19] models of Huggingface-transformers are used to implement the pretrained model ELECTRA, ERNIE, and ALBERT, respectively.

Evaluation metric

Artificial intelligence applications require large quantities of training and test data. This demand presents significant challenges not only concerning the availability of such data, but also regarding its quality. Incomplete, erroneous or inappropriate training data can lead to unreliable models that produce ultimately poor decisions (Budach et al., 2022). Therefore, a comprehensive and rigorous data quality assessment is important for dataset construction. Three quality attributes are comprehensiveness, correctness, and variety, which are most critical to "fit for purpose" of deep learning (Chen, Chen & Ding, 2021). We use qualitative or quantitative methods to evaluate the quality of datasets produced by our dataset construction method from the

[7]https://api.fanyi.baidu.com
[8]https://pypi.org/project/jieba
[9]https://pypi.org/project/gensim
[10]https://pypi.org/project/sklearn
[11]https://huggingface.co/bert-base-uncased/tree/main
[12]https://github.com/zhoujx4/NLP-Series-sentence-embeddings
[13]It will obtain tokens, which is the basic unit in which a computer processes text.
[14]https://huggingface.co/roberta-base/tree/main
[15]https://huggingface.co/mbart-large-cc25/tree/main
[16]https://huggingface.co
[17]https://huggingface.co/electra-base-discriminator/tree/main
[18]https://huggingface.co/ernie-2.0-base-en/tree/main
[19]https://huggingface.co/albert-base-v2/tree/main

perspective of such three quality attributes. According to the data quality assessment framework proposed by Chen, Chen & Ding (2021), we make the qualitative evaluation of the comprehensiveness of the dataset by checking the data source, the qualitative evaluation of the correctness of the dataset by manually checking samples and the quantitative evaluation of the variety of the dataset by checking the uniqueness of samples, and checking the overlap of train, valid and test sets. In addition, according to the conclusion made by Chen, Pieptea & Ding (2022), we design a group of experiments directly for CLS to quantitatively evaluate the effect of TAPT and the quality of data obtained by it.

In the experiments to verify En2Zh_Sum and TAPT, we use ROUGE (Lin, 2004) to evaluate CLS results, specifically using rouge-metric[20] library. Note that the standard ROUGE metric only evaluates English *summaries*, and thus a special treatment is applied to evaluate Chinese *summaries* in our study, i.e., the *summaries* are segmented by character granularity and then spliced with space characters.

In the experiment to select the most appropriate pretrained model, we use the average accuracy of predicted words equal to the masked words to measure the predictive power of the model.

Experimental Results and Analysis

Evaluation of dataset quality

Check of the comprehensiveness

One way of the evaluation is to evaluate the data collection procedure and data sources (Chen, Chen & Ding, 2021). The process of our dataset construction method is shown in Fig 1. Firstly, we use MSF to remove low-quality samples from the data source, ensuring quality at the beginning of the construction. Then, we use the excellent Baidu Translation service to translate the *text* in the data source from Chinese to English, ensuring the quality of the collection procedure. Finally, we use TAPT to expand the CLS dataset obtained in the previous step, which increases the data size while ensuring the sample quality. We select the LCSTS (Hu, Chen & Zhu F, 2015) dataset as the data source. LCSTS is a benchmark dataset of ATS obtained from Sina Weibo. Its texts are short and noisy, which not only makes the model easier to learn from data, but also increases the generalization performance to a certain extent. The authors manually mark the correlation between the *text* and the *summary*. This correlation reflects quality of samples. We can select samples with different correlation scores according to specific tasks, so as to obtain the valid set and test set of appropriate quality. The above analysis shows that En2Zh_Sum is of good comprehensiveness and reliable quality.

Check of the correctness

The most straightforward way to check the correctness of a dataset is to check the sample data manually (Chen, Chen & Ding, 2021). We randomly sample 100 samples from the train set, valid set and test set of En2Zh_Sum, respectively, and check them manually. Three graduate students are asked to check each sample from three independent perspectives: (1) correlation, (2)

[20]https://pypi.org/project/rouge-metric

conciseness, and (3) fluency. Each perspective is assessed with a score ranging from 1 (worst) to 5 (best). Table 7 presents the average results.

As shown in Table 7, no matter which dataset, *summaries* and their corresponding *texts* have well conciseness and fluency. In $\text{LCSTS}_{\text{MSF}}$ and En2Zh_Sum, *summaries* can well reflect the content of their corresponding *texts*. However, in LCSTS, the correlation between *summaries* and their corresponding *texts* is obviously low. It shows that En2Zh_Sum is of good correctness and reliable quality. The increase of the score of correlation from LCSTS to $\text{LCSTS}_{\text{MSF}}$ indicates the effect of MSF on improving the quality of MS data set.

Check of the variety

Some properties of variety need to be checked are the unique data items in a dataset and the overlap in train, valid and test sets (Chen, Chen & Ding, 2021). We calculate the uniqueness ratio of the train, valid and test sets of En2Zh_Sum respectively, as well as the overlap ratio among them. Table 8 presents the checking results.

As shown in Table 8, samples in En2Zh_Sum are unique, and there is no overlap among the three splits. It shows that En2Zh_Sum is of good variety and reliable quality.

Experimental evaluation

The experimental study in machine learning and deep learning can quantitatively evaluate the quality of the dataset (Chen, Pieptea & Ding, 2022). We fine-tune MBART on the augmented train subsets of NCLS and compare the performance with the results of many CLS studies on the full train set. The experimental results are listed in Table 9.

The experimental results show that the direct application of MBART does not perform well for either English-Chinese or Chinese-English CLS, which suggests that even if the pretrained model has strong performance, it cannot be directly applied to CLS without learning from specific data. MBART_{ft}-CLS (the MBART fine-tuned on the train subset) achieves **+18.77** ROUGE-1, **+13.2** ROUGE-2, **+15.84** ROUGE-L for English-Chinese CLS and **+1.42** ROUGE-1, **+0.11** ROUGE-2, **+4.98** ROUGE-L for Chinese-English CLS compared to the state-of-the-art performance, which shows that the pretrained model can significantly improve the performance of CLS system. $\text{(MBART+TPTA)}_{\text{ft}}$-CLS (the MBART fine-tuned on the augmented train subset) achieve **+19.83** ROUGE-1, **+15.4** ROUGE-2, **+17.4** ROUGE-L for English-Chinese CLS and **+1.49** ROUGE-1, **+0.31** ROUGE-2, **+4.99** ROUGE-L for Chinese-English CLS compared to the state-of-the-art performance, which shows that TAPT can generate high-quality CLS samples and improve CLS performance, and indirectly validates the quality of En2Zh_Sum.

We can see that after fine-tuning the CLS task on the MBART, performance is well above the baselines. The difficulty of improving performance again at this point is enormous. The essence of data augmentation to improve performance is to increase samples of train set. MBART_{ft}-CLS has learned the train set well, while $\text{(MBART+TPTA)}_{\text{ft}}$-CLS only has more training samples than MBART_{ft}-CLS. So $\text{(MBART+TPTA)}_{\text{ft}}$-CLS won't have a significant performance improvement over MBART_{ft}-CLS, but it is a satisfying and surprising result that the performance improvement of over is about 1% (English-Chinese) and 0.1% (Chinese-English). The bi-direction performance has a big difference. There are two main reasons: (1) MBART is a multilingual pretrained model. Due to the differences in the pre-training corpus and the characteristics of Chinese and English, the language ability of the model is different. Therefore, this model can be regarded as two different models when conducting CLS experiments in two different cross-lingual directions. (2) The datasets for CLS experiments in bi-direction are different. The dataset used for English-Chinese CLS is En2ZhSum, and the dataset used for Chinese-English CLS is Zh2EnSum. The statistics are shown in Table 5. Their source, size, length of samples and other aspects have obvious differences. To sum up, it is quite normal for two different pretrained models to have big differences in experimental results on different datasets.

The size of En2Zh_Sum is shown in Table 10. To validate the quality of En2Zh_Sum simply and intuitively, we randomly sample the one-seventh of train set (400,000 samples) to fine-tune MBART and test on whole test set. The result is shown in Table 11. It shows that the CLS model can achieve good performance with only part of En2Zh_Sum, which proves that our dataset En2Zh_Sum is of high quality and the effectiveness and feasibility of our dataset construction method of CLS.

Choice of the pretrained model

We randomly sample five English texts from NCLS, and randomly select ten words from each text, as shown in Table 12. And we use five pre-trained models of BERT, ELECTRA, ERNIE, RoBERTa and ALBERT to predict the masked tokens. The average prediction accuracy is shown in Table 13.

The experimental results show that RoBERTa has the highest accuracy, which indicates that it has the optimal performance for predicting words. Table 14 shows some typical results of applying RoBERTa in TAPT. The result of the first text is the same as the original text, and the result of the second text is slightly different from the original text, which shows that RoBERTa can ensure both similarities and differences between the generated text and the original text to generate suitable new samples for augmentation.

One confusing result is that the performance of ERNIE is 0. Table 13 shows the average accuracy of predicted words equal to the masked words to measure the predictive power of the model. The average accuracy is the mean of the ratio of the number of predicted words equal to the mask words to the total number of mask words in all experimental samples. The real result of the experiment is that ERNIE don't get a single word right, so the average accuracy is 0. ERNIE is a very powerful pretrained model right, which improves MLM of BERT. Although the performance of ERNIE on various NLP tasks is greatly improved, the experimental result shows that its ability to predict words directly actually decreased, which is unsuitable for TAPT.

Conclusions

In this paper, we propose a dataset construction method of CLS that jointly supervises quality and scale, and build a high-quality and large-scale English-Chinese CLS dataset En2Zh_Sum. Our method uses MSF to remove low-quality MS samples from the perspectives of character and semantics to supervise quality, and TAPT which uses self-attention and MLM to increase samples to supervise scale. The experimental results show that our method can not only filter out low-quality samples comprehensively but also augment data scale flexibly and effectively to obtain a high-quality and large-scale CLS dataset at a lower cost.

Currently, there are few methods to evaluate and improve the quality of MS datasets. MSF is the first method to improve the quality of MS datasets by measuring the degree to which the *summary* reflects the content of its original *text* from the perspectives of character and semantics. It is simple and effective, and can be generalized to handle similar types of non-parallel text pairs. Compared with existing text augmentation algorithms based on pretrained models, TAPT utilizes self-attention to more rationally select words to be replaced. In the dynamic synonym replacement, TAPT uses a more powerful pre-training model to get the best performance of predictive words. TAPT encourages researchers to make reasonable use of the features of pretrained models, and can be used to augment texts for other tasks. Our dataset construction method is the first systematic method to build CLS datasets. In the process of construction, effective techniques are adopted to strictly supervise the quality and scale. It can be used to build more CLS datasets. The datasets constructed by our method can be directly used for future research.

In future work, we will follow the ideas of our method to optimize the supervision process of quality and scale. In terms of quality supervision, we intend to measure more accurately how well the *summary* reflects the content of the original *text* from the perspective of semantics. In terms of scale supervision, we will consider how best to leverage the capabilities of the pretrained model to expand samples with higher quality.

Acknowledgements

We thank reviewers for their helpful comments and editage (www.editage.cn) for its linguistic assistance during the preparation of this manuscript.

References

[1] Anaby-Tavor A, Carmeli B, Goldbraich E, Kantor A, Kour G, Shlomov S, Tepper N, Zwerdling N. 2019. Not enough data? Deep learning to the rescue!. arXiv preprint arXiv:1911.03118

[2] Ayana, Shen S, Chen Y, Yang C, Liu Z, Sun M. 2018. Zero-shot cross-lingual neural headline generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* 26(12):2319-2327

[3] Bai Y, Gao Y, Huang H. 2021. Cross-lingual abstractive summarization with limited parallel resources.In: the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing (ACL-IJCNLP). 6910-6924

[4] Bai Y, Huang H, Fan K, Gao Y, Chi Z, Chen B. 2021. Bridging the gap: cross-lingual summarization with compression rate. arXiv preprint arXiv:2110.07936

[5] Budach L, Feuerpfeil M, Ihde N, Nathansen A, Noack N, Patzlaff H, Harmouch H, Naumann F. 2022. The Effects of Data Quality on Machine Learning Performance. arXiv preprint arXiv:2207.14529v4

[6] Cao Y, Liu H, Wan X. 2020. Jointly learning to align and summarize for neural cross-lingual summarization. In: the 58th Annual Meeting of the Association for Computational Linguistics (ACL). 6220-6231

[7] Chen H, Chen J, Ding J. 2021. Data Evaluation and Enhancement for Quality Improvement of

Machine Learning. IEEE Transactions on Reliability 70(2):831-847
[8] Chen H, Pieptea L, Ding J. 2022. Construction and Evaluation of a High-Quality Corpus for Legal Intelligence Using Semiautomated Approaches. IEEE Transactions on Reliability 71(2):657-673
[9] Chi Z, Dong L, Wei F, Wang W, Mao X, Huang H. 2020. Cross-lingual natural language generation via pre-training. In: the AAAI Conference on Artificial Intelligence (AAAI). 7570-7577
[10] Clark K, Luong M, Le Q, Manning C. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In: International Conference on Learning Representations (ICLR).
[11] Devlin J, Chang M, Lee K, Toutanova K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. In: the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL). 4171-4186
[12] Dou Z, Kumar S, Tsvetkov Y. 2020. A deep reinforced model for zero-Shot cross-lingual summarization with bilingual semantic similarity Rewards. In: 4th Workshop on Neural Generation and Translation. 60-68
[13] Duan X, Yin M, Zhang M, Chen B, Luo W. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In: the 57th Annual Meeting of the Association for Computational Linguistics (ACL). 3162-3172
[14] Erkan G, Radev D. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)* 457-479
[15] Fatima M, Strube M. 2021. A novel wikipedia based dataset for monolingual and cross-lingual summarization. In: the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). 39-50
[16] Hermann K, Kočiský T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P. 2015. Teaching machines to read and comprehend. In: the 28th International Conference on Neural Information Processing Systems (NIPS). 1693-1701
[17] Hou Y, Liu Y, Che W, Liu T. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In: the 27th International Conference on Computational Linguistics (COLING). 234-1245
[18] Hu B, Chen Q, Zhu F. 2015. LCSTS: a large scale Chinese short text summarization dataset. In: the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). 1967-1972
[19] Kobayashi S. 2018. Contextual Augmentation: Data augmentation by words with paradigmatic relations. In: the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL). 452-457
[20] Ladhak F, Durmus E, Cardie C, Mckeown K. 2020. WikiLingua: a new benchmark dataset for cross-lingual abstractive summarization. In: the Findings of the Association for Computational Linguistics: EMNLP 2020. 4034-4048
[21] Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In: International Conference on Learning Representations (ICLR).
[22] Leuski A, Lin C, Zhou L, Germann U, Och F, Hovy E. 2003. Cross-lingual c*st*rd: English access to Hindi information. *ACM Transactions on Asian Language Information Processing (TALIP)* 2(3):245-269
[23] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: the 58th Annual Meeting of the Association for Computational Linguistics (ACL). 7871-7880
[24] Li B, Zhou H, He J, Wang M, Yang Y, Li L. 2020. On the sentence embeddings from pre-trained language models. In: 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 9119-9130
[25] Lin C. 2004. ROUGE: a package for automatic evaluation of summaries. In: Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL. 74-81

[26] Liu Y, Gu J, Goyal N, Li X, Edunov S, Ghazvininejad M, Lewis M, Zettlemoyer L. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics (TACL)* 8:726-742
[27] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692
[28] Loshchilov I, Hutter F. 2019. Fixing weight decay regularization in adam. In: International Conference on Learning Representations (ICLR).
[29] Macqueen J. 1966. Some Methods for Classification and Analysis of Multi Variate Observations. In: Berkeley Symposium on Mathematical Statistics and Probability. 281-297
[30] Mikolov T, Chen K, Corrado G, Dean J. 2013. Efficient estimation of word representations in vector space. In: the 1st International Conference on Learning Representations, ICLR 2013-Workshop Track Proceedings.
[31] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. 2013. Distributed representations of words and phrases and their compositionality. In: the 26th International Conference on Neural Information Processing Systems (NIPS). 3111-3119
[32] Napoles C, Gormley M, Durme B. 2012. Annotated Gigaword. In: the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction. 95-100
[33] Nguyen K, Daumé H. 2019. Global voices: crossing borders in automatic news summarization. In: the 2nd Workshop on New Frontiers in Summarization. 90-97
[34] Orăsan C and Chiorean O. 2008. Evaluation of a cross-lingual Romanian-English multi-document summarizer. In: Language Resources and Evaluation Conference (LREC).
[35] Ouyang J, Song B, McKeown K. 2019. A robust abstractive system for cross-lingual summarization. In: the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). 2025-2031
[36] Over P, Dang H, Harman D. 2007. DUC in context. *Information Processing and Management: an International Journal* 43(6):1506-1520
[37] Radford A, Narasimhan K, Salimans T, Sutskever I. 2018. Improving language understanding by generative pre-training. URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper.pdf.
[38] Schwartz E, Karlinsky L, Shtok J, Harary S, Marder M, Feris R, Kumar A, Giryes R, Bronstein A. 2018. Delta-encoder: An effective sample synthesis method for few-shot object recognition. In: the 32nd International Conference on Neural Information Processing Systems (NIPS). 2850-2860
[39] Sennrich R, Haddow B, Birch A. 2016. Improving neural machine translation models with monolingual data. In: the 54th Annual Meeting of the Association for Computational Linguistics (ACL). 86-96
[40] Sennrich R, Haddow B, Birch A. 2016. Neural machine translation of rare words with subword units. In: the 54th Annual Meeting of the Association for Computational Linguistics (ACL). 1715-1725
[41] Siddharthan A, McKeown K. 2005. Improving multilingual summarization: using redundancy in the input to correct MT errors. In: HLT/EMNLP-2005. 33-40
[42] Su J, Cao J, Liu W, Ou Y. 2021. Whitening sentence representations for better semantics and faster retrieval. arXiv preprint arXiv:2103.15316
[43] Sun Y, Wang S, Li Y, Feng S, Tian H, Wu H, Wang H. 2020. ERNIE 2.0: A continual pre-training framework for language understanding. In: AAAI Conference on Artificial Intelligence (AAAI). 8968-8975
[44] Takase S, Okazaki N. 2020. Multi-task learning for cross-lingual abstractive summarization. arXiv preprint arXiv:2010.07503
[45] Tang Z, Xiao Q, Zhu L, Li K, Li K. 2019. A semantic textual similarity measurement model based on the syntactic-semantic representation. *Intelligent Data Analysis* 23(4):933-950,
[46] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Kaiser Ł, Polosukhin I. 2017. Attention is all you need. In: the 31st International Conference on Neural Information Processing

Systems (NIPS). 5998-6008
[47] Wan X, Li H, Xiao J. 2010. Cross-language document summarization based on machine translation quality prediction. In: the 48th Annual Meeting of the Association for Computational Linguistics (ACL). 917-926
[48] Wan X, Luo F, Sun X, Huang S, Yao J. 2019. Cross-language document summarization via extraction and ranking of multiple summaries. *Knowledge and Information Systems (KAIS)* 58(2):481-499
[49] Wan X. 2011. Using bilingual information for cross-language document summarization. In: the 49th Annual Meeting of the Association for Computational Linguistics (ACL). 1546-1555
[50] Wang J, Zhang Y, Yu Z, Huang Y. 2021. Semi-supervised adversarial Chinese-Vietnamese cross-lingual summarization generation method using word alignment. *Journal of Chinese Computer Systems* 1-8 (in Chinese)
[51] Wei J, Zou K. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In: the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 6382-6388
[52] Wu X, Lv S, Zang L, Han J, Hu S. 2019. Conditional BERT contextual augmentation. In: International Conference on Computational Science. 84-95
[53] Xie Q, Dai Z, Hovy E, Luong M, Le Q. 2019. Unsupervised data augmentation for consistency training. arXiv preprint arXiv: 1904.12848
[54] Xu R, Zhu C, Shi Y, Zeng M, Huang X. 2020. Mixed-lingual pre-training for cross-lingual summarization. In: the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (AACL-IJCNLP). 536-541
[55] Yao J, Wan X, Xiao J. 2015. Phrase-based compressive cross-language summarization. In: the 2015 Conf on Empirical Methods in Natural Language Processing (EMNLP). 118-127
[56] Yin M, Shi X, Yu H, Duan X. 2020. Cross-lingual sentence summarization system based on contrastive attention mechanism . *Computer Engineering* 46(5):86-93 (in Chinese)
[57] Yu A, Dohan D, Luong M, Zhao R, Chen K, Norouzi M, Le Q. 2018. QANet: Combining local convolution with global self-attention for reading comprehension. In: International Conference on Learning Representations (ICLR)
[58] Zhang J, Zhou Y, Zong C. 2016. Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* 24(10):1842-1853
[59] Zhu J, Wang Q, Wang Y, Zhou Y, Zhang J, Wang S, Zong C. 2019. NCLS: neural cross-lingual summarization. In: the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 3054-3064
[60] Zhu J, Zhou Y, Zhang J, Zong C. 2020. Attend, translate and summarize: an efficient method for neural cross-lingual summarization. In: the 58th Annual Meeting of the Association for Computational Linguistics (ACL). 1309-1321

Figure 1

The process of the collection method

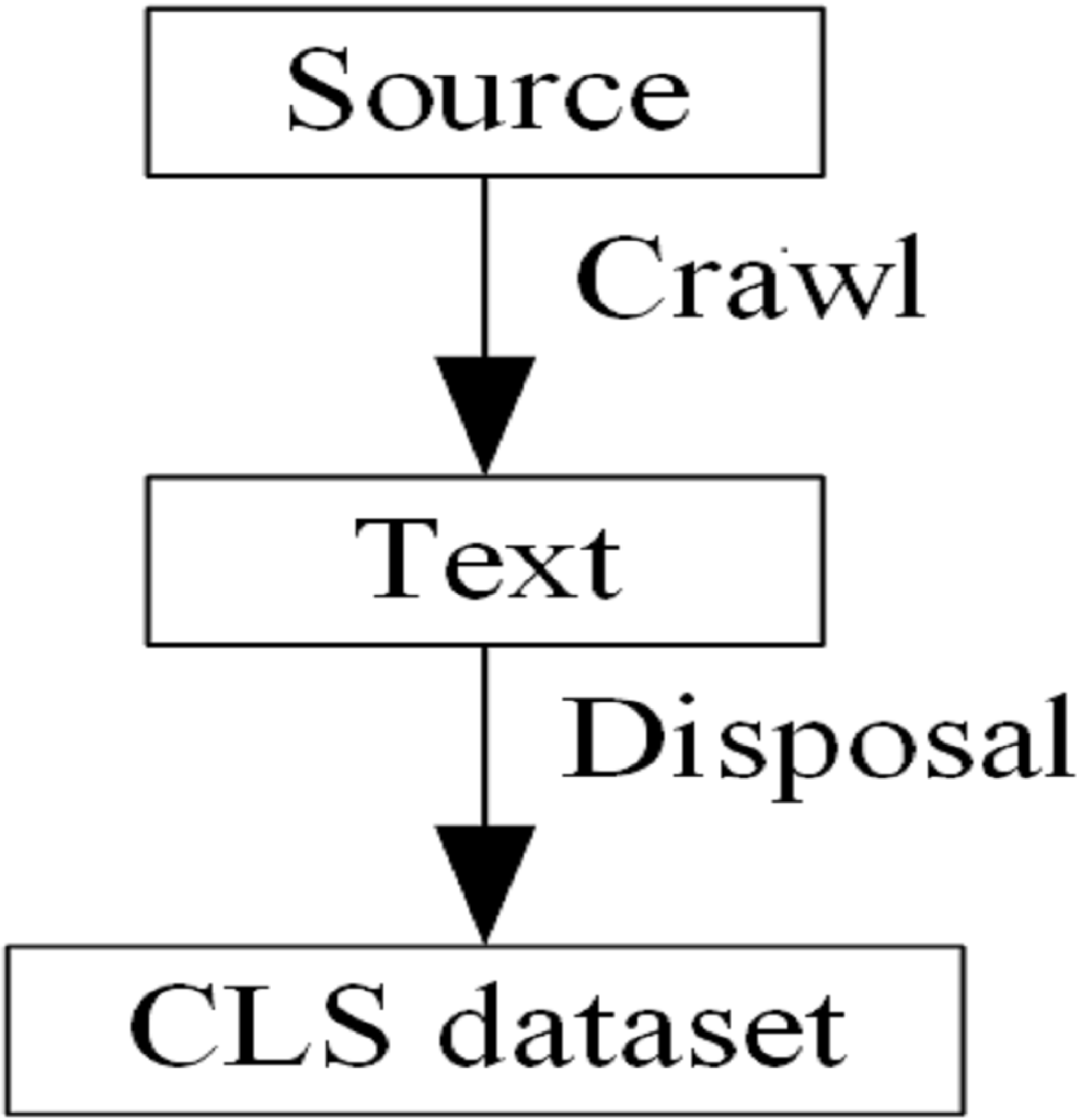

Figure 2

The process of the transformation method

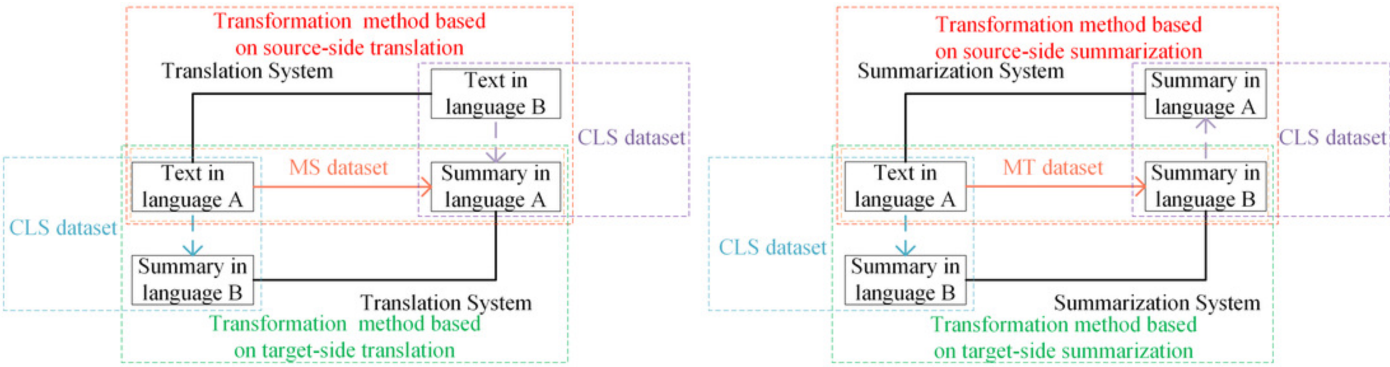

Figure 3

The process of the proposed dataset construction method of CLS

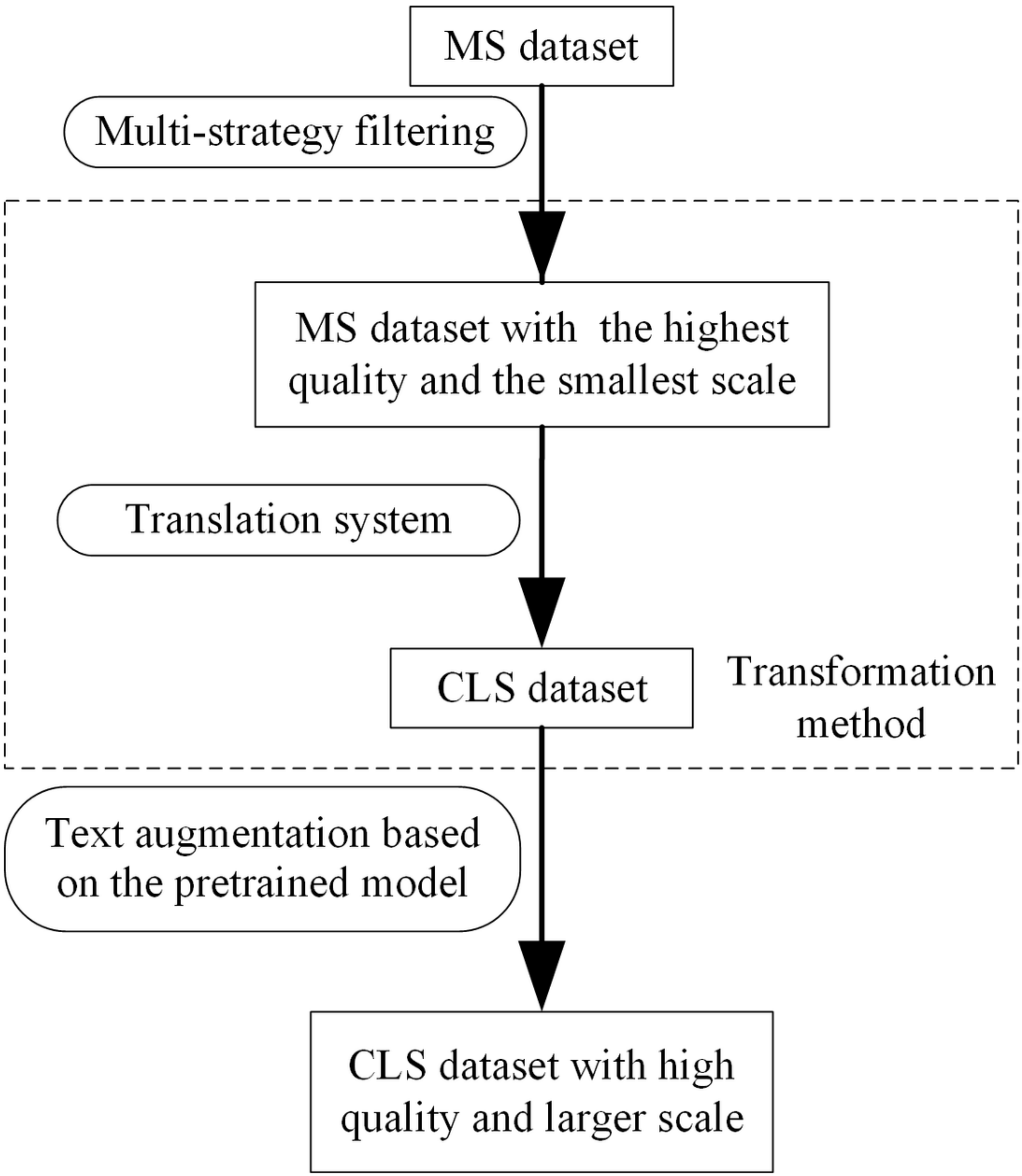
MS dataset
Multi-strategy filtering
MS dataset with the highest quality and the smallest scale
Translation system
CLS dataset
Transformation method
Text augmentation based on the pretrained model
CLS dataset with high quality and larger scale

Figure 4

The overall process of MSF

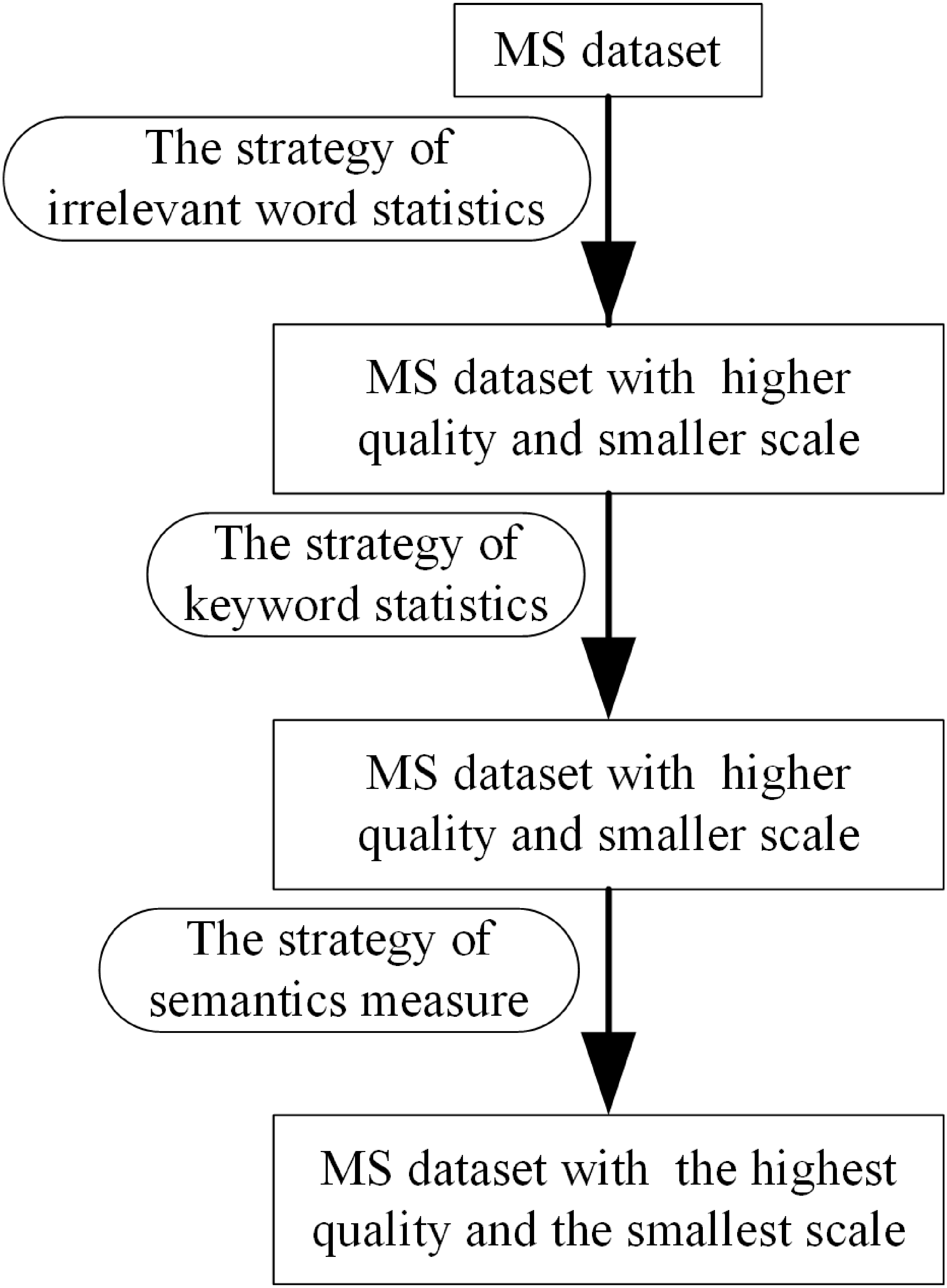
MS dataset
The strategy of irrelevant word statistics
MS dataset with higher quality and smaller scale
The strategy of keyword statistics
MS dataset with higher quality and smaller scale
The strategy of semantics measure
MS dataset with the highest quality and the smallest scale

Figure 5

The process of TAPT

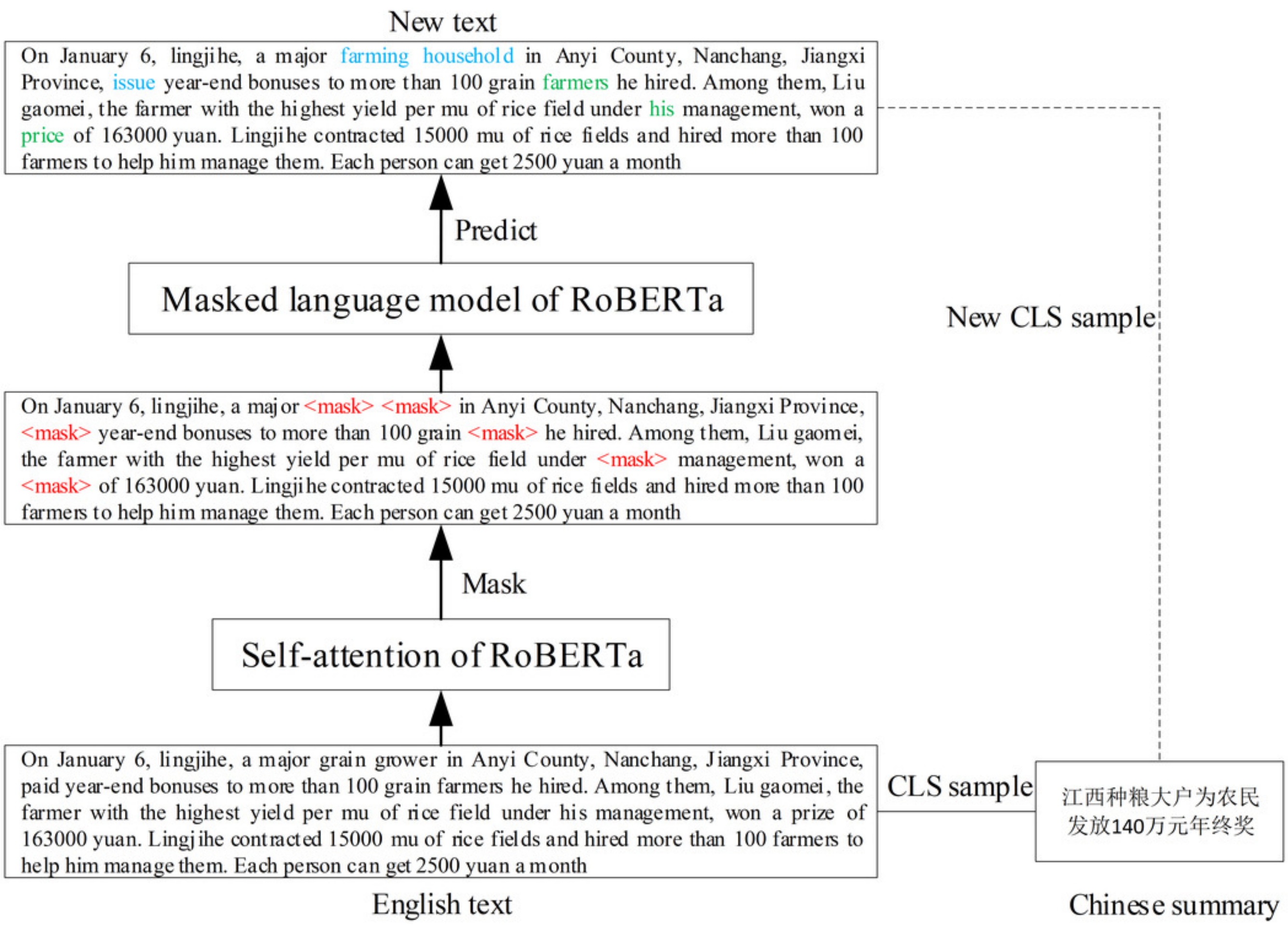

Table 1(on next page)

Samples of the LCSTS dataset.

The straight underline denotes keywords that appear in both the text and the summary. The red text denotes content that appears in the summary but not in the text and is unrelated to the text. The wavy underline denotes content that appears in the summary but not in the text and reflects key information.

LCSTS	
Text	Reference summary
近日国家能源局公布了《可再生能源发电并网驻点甘肃监管报告》，报告是在国家能源局对甘肃进行3个月可再生能源发电监管之后形成的。《报告》显示甘肃省可再生能源发电并网存在诸多问题。	能源局监管甘肃可再生能源全省弃风率超20%。
一辆小轿车，一名女司机，竟造成9死24伤。日前，深圳市交警局对事故进行通报：从目前证据看，事故系司机超速行驶且操作不当导致。目前24名伤员已有6名治愈出院，其余正接受治疗，预计事故赔偿费或超一千万元。	深圳机场9死24伤续：司机全责赔偿或超千万。
中国有句古话“养儿防老”，而这三十年来所执行的强制计划生育政策使得“养儿防老”变为了不可能，绝大多数成员的养老问题除了依靠社会力量之外别无他路。养老不光是老人们所面临的问题，老无所依使得未老的社会成员也开始惶恐不安。	俞天任：老龄化问题不解决将亡族灭种。

1

Table 2 (on next page)

An overview of CLS datasets.

*The dataset contains many sub-datasets with cross-lingual directions. The average size of all sub-datasets is used to represent the size of this dataset.

Dataset	Method Type	Mode	Scale	Open Source
Ladhak et al. (2020)	Collection	Auto+Manual	18k*	All
Nguyen & Daumé (2019)	Collection	Auto+Manual	gv-snippet: 1k* gv-crowd: 0.2k*	All
Fatima & Strube (2021)	Collection	Auto+Manual	W-CLS: 51k S-CLS: 48k	All
Ayana et al. (2018)	Transformation	Auto	3.8M	Not
Duan et al. (2019)	Transformation	Auto	3.8M	Some
Zhu et al. (2019)	Transformation	Auto	En2ZhSum: 371k Zh2EnSum: 1.7M	All

Table 3 (on next page)

An overview of text augmentation algorithms.

Algorithm	Object	Model	Method
Wei & Zou (2019)	Word	-	synonym replacement, random insertion, random exchange, random deletion
Kobayashi (2018)	Word	Bidirectional Language Model	synonym replacement
Wu et al. (2019)	Word	BERT	synonym replacement
Yu et al. (2018)	Text	-	back-translation
Xie et al. (2019)	Text	-	back-translation
Hou et al. (2018)	Text	Seq2Seq Model	generate new texts
Anaby-Tavor et al. (2019)	Text	GPT-2	generate new texts

Table 4(on next page)

Workflow of Whitening-h.

Algorithm 1	Whitening-h

Input: Existing embeddings $\{\boldsymbol{z}_k'\}_{k=1}^{2N}$ and reserved dimensionality h

1: compute Mean μ and variance Σ of $\{\boldsymbol{z}_k'\}_{k=1}^{2N}$
2: compute U , Λ and $U^T = SVD(\Sigma)$
3: compute $W = \left(U\sqrt{\Lambda^{-1}}\right)[:, :h]$
4: **for** $k = 1, 2, \dots, 2N$ **do**
5: $\quad \tilde{\boldsymbol{z}}_k' = (\boldsymbol{z}_k' - \mu)W$
6: **end for**

Output: Transformed embeddings $\{\tilde{\boldsymbol{z}}_k'\}_{k=1}^{2N}$

Table 5(on next page)

Statistics on the NCLS dataset.

[1]Num denotes the size of the dataset. [2]SrcAvgToken denotes the average token number of source language texts. [3]SrcMaxToken denotes the maximal token number of source language texts. [4]TgtAvgToken denotes the average token number of target language summaries. [5]TgtMaxToken denotes the maximal token number of target language summaries.

En2ZhSum	Train	Valid	Test	Zh2EnSum	Train	Valid	Test
Num[1]	364,687	3,000	3,000	Num[1]	1,693,713	3,000	3,000
SrcAvgToken[2]	942.7	949.1	930.2	SrcAvgToken[2]	73.4	73.3	73.6
SrcMaxToken[3]	12,498	7,547	8,635	SrcMaxToken[3]	134	113	119
TgtAvgToken[4]	70.0	70.1	69.9	TgtAvgToken[4]	20.6	20.6	21.5
TgtMaxToken[5]	593	242	260	TgtMaxToken[5]	70	48	53

Table 6(on next page)

Key parameters of experiments.

[1]Tokenizer denotes the tokenize algorithm. [2]En2Zh I/O length denotes the input/output sequence length of model in English-to-Chinese CLS. [3]Zh2En I/O length denotes input/output sequence length of the model in Chinese-to-English CLS. [4]Iter denotes the iterations at the end of fine-tuning.

Parameter	Setup
CLS Tokenizer[1]	BPE
En2Zh I/O length[2]	550/100
Zh2En I/O length[3]	80/60
Iter[4]	100,000

Table 7 (on next page)

Human evaluation results on the three datasets.

[1]CR, CC, and FL denote the scores for correlation, conciseness, and fluency, respectively. $\text{LCSTS}_{\text{MSF}}$ represents the samples left after MSF is used on the LCSTS dataset.

Dataset	**Role**	**Split**	**CR**[1]	**CC**[1]	**FL**[1]
LCSTS	Source	Train	3.48	3.80	4.08
		Valid	3.56	3.79	4.01
		Test	3.62	3.83	4.03
LCSTS$_{MSF}$	Intermediate	Train	4.10	3.77	4.05
		Valid	4.05	3.84	4.09
		Test	4.09	3.81	4.02
En2Zh_Sum	Final	Train	4.08	3.78	4.12
		Valid	4.12	3.86	4.04
		Test	4.06	3.82	4.02

Table 8 (on next page)

Checking results of the uniqueness and overlap of En2Zh_Sum splits.

Split	Uniqueness Ratio	Overlap Ratio
Train	100%	0% (with Valid)
Valid	100%	0% (with Test)
Test	100%	0% (with Train)

Table 9(on next page)

The results of CLS experiments.

ROUGE F1 scores (%) on En2ZhSum and Zh2EnSum test sets. † denotes the previous best performance. * denotes the results of fine-tuning MBART on the train subsets. The bold number denotes the results of fine-tuning MBART on the augmented train subsets.

Method	English-to-Chinese CLS			Chinese-to-English CLS		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
-Pipeline methods-						
TETran	26.15	10.60	23.24	23.09	7.33	18.74
TLTran	30.22	12.20	27.04	33.92	15.81	29.86
GETran	28.19	11.40	25.77	24.34	9.14	20.13
GLTran	32.17	13.85	29.43	35.45	16.86	31.28
-End-to-end methods-						
NCLS	36.82	18.72	33.20	38.85	21.93	35.05
NCLS-MT	40.23	22.32	36.59	40.25	22.58	36.21
NCLS-MS	38.25	20.20	4.76	40.34	22.65	36.39
XNLG-CLS	39.85	24.47	28.28	38.34	19.65	33.66
ATS	40.47	22.21	36.89	40.68	24.12†	36.97
MLPT	43.50†	25.41†	29.66	41.62†	23.35	37.26†
RL-XSIM	42.83	23.30	39.29†	-	-	-
MCLAS	42.27	24.60	30.09	35.65	16.97	31.14
CSC	-	-	-	40.30	21.43	35.46
-The proposed method-						
MBART-CLS	14.59	4.31	10.87	0.71	0.04	0.70
MBART$_{ft}$-CLS	62.27*	38.61*	55.13*	43.04*	24.23*	42.24*
(MBART+PTA)$_{ft}$-CLS	**63.33**	**40.81**	**56.69**	**43.11**	**24.43**	**42.25**

Table 10 (on next page)

Data size of the En2Zh_Sum.

En2Zh_Sum	Train	Valid	Test
Size	2,810,266	10,000	10,000

Table 11(on next page)

ROUGE F1 scores (%) on the En2Zh_Sum test set.

Model	English-Chinese CLS		
	ROUGE-1	ROUGE-2	ROUGE-L
$MBART_{ft}$-CLS	46.30	23.80	42.45

Table 12 (on next page)

The experimental data.

[MASK] indicates that the token at this position is masked.

Text	Masked token
According to [MASK] latest Reuters news, the U.S. police updated the number of casualties in the Denver shooting [MASK] 12 deaths and 58 injuries. On Friday night local time, 30 [MASK] people were [MASK] hospitalized for treatment, [MASK] of whom were in [MASK] condition. [MASK] 24-year-old [MASK] James Egan Holmes is being interrogated and [MASK] motive has not [MASK] determined yet. Compiled and reported by CNTV Jiang Yiyi.	'the', 'as', 'injured', 'still', '11', 'critical', 'The', 'suspect', 'his', 'been'
Robin Lee, member of [MASK] CPPCC National Committee [MASK] CEO [MASK] Baidu, [MASK] that his proposal this [MASK] mainly [MASK] on using the Internet to improve the current network registration system. He [MASK] that the restrictions on commercial institutions to [MASK] out online registration business in some [MASK] should be lifted, and the allocation of medical [MASK] should be optimized with the help of social forces	'the', 'and', 'of', 'revealed', 'year', 'focused', 'suggested', 'carry', 'regions', 'resources'
According [MASK] the news on the 21st, the continuous rainstorm caused [MASK] torrents at k806 + 500 of national highway [MASK] in Guangyuan, Sichuan, and some roads were damaged. At present, it is impossible to predict the opening time. At about 6:00 on the 21st, flash floods [MASK] out at Tashan Bay on national highway 212, [MASK] about [MASK] meters of asphalt concrete subgrade was washed away, [MASK] local uplift [MASK] the pavement and subsidence of the [MASK] Edited and [MASK] by CCTV yanghanning.	'to', 'mountain', '212', 'broke', 'and', '600', 'with', 'of', 'subgrade.', 'reported'
From now on, the Municipal Bureau of urban and rural planning [MASK] launched [MASK] overall conceptual planning solicitation activity [MASK] 15 xiangjiangzhou islands. The overall conceptual planning solicitation of xiangjiangzhou Island [MASK] two [MASK] at the same time, [MASK] the International Solicitation [MASK] world-class professional design units and the solicitation for [MASK] schemes" for the public. For details, please visit the official website of the Municipal Bureau of [MASK] and rural [MASK]	'has', 'an', 'for', 'opened', '"channels"', 'namely,', 'for', '"good', 'urban', 'planning.'
Liang [MASK] a lawyer from Zhonglun law [MASK] suggested that female [MASK] should [MASK] the police at the first time. As for the [MASK] of applying glue to long hair, which is [MASK] infringement [MASK] physical rights in civil law, although it is bad, it has not risen to the level of crime in [MASK] It can only be imposed with administrative penalties [MASK] as fines and criticism and education in accordance with [MASK] law on public security administration and punishment.	'Jing,', 'firm,', 'victims', 'call', 'act', 'an', 'of', 'law.', 'such', 'the'

Table 13 (on next page)

The average accuracy of predictions.

Model	Accuracy
BERT	0.44
ELECTRA	0.42
ERNIE	0
RoBERTA	0.5
ALBERT	0.24

1

Table 14 (on next page)

Results of the RoBERTa-based TAPT.

Red words denote the masked words. Green words denote the same prediction result as the original words. Blue words denote a different prediction result from the original words.

Original text	Generated text
By the end of last year, the balance of broad money (M2) in China had reached 97.42 trillion yuan, and there was no doubt that it would exceed one billion yuan. This figure is 1.5 times that of the United States, 4.9 times that of Britain and 1.7 times that of Japan. This figure is close to a quarter of the total global money supply. It is no exaggeration to say that China has become the largest country in the global money stock	By the end of last year, the balance of broad money (M2) in China had reached 97.42 trillion yuan, and there was no doubt that it would exceed one billion yuan. This figure is 1.5 times that of the United States, 4.9 times that of Britain and 1.7 times that of Japan. This figure is close to a quarter of the total global money supply. It is no exaggeration to say that China has become the largest country in the global money stock
It was learned from authoritative sources yesterday that Zhong\'an online property insurance company, jointly established by Alibaba\'s Jack Ma, Ping An\'s Jack Ma and Tencent\'s Jack Ma, has now completed the regulatory approval process. It is expected that the CIRC will officially issue an approval document approving its preparation soon. It is reported that Eurasia Ping, a mysterious rich businessman, will take the post of chairman, which is jointly recommended by the "three horses" "	It was learned from authoritative sources yesterday that Zhong\'an online property insurance company, jointly established by Alibaba\'s Jack Ma, Ping An\'s Jack Ma and Tencent\'s Jack Ma, has now completed the regulatory approval process. It is expected that the CIRC will officially issue an official document approving its preparation soon. It is reported that Eurasia Ping, a mysterious rich businessman, will take the role of chairman, which is jointly recommended by the "three horses" "