

Automatic Generation of Artificial Space Weather Forecast Product Based on Sequence-to-sequence Model*

LUO Guanting^{1,2,3} ZOU Yenan^{1,3} CAI Yanxia^{1,3}

1(State Key Laboratory of Space Weather, National Space Science Center, Chinese Academy of Sciences, Beijing 100190)

2(University of Chinese Academy of Sciences, Beijing 100049)

3(Key Laboratory of Science and Technology on Environment Space Situation Awareness, Chinese Academy of Sciences, Beijing 100190)

Abstract Both analyzing a large amount of space weather observed data and alleviating personal experience bias are significant challenges in generating artificial space weather forecast products. With the use of natural language generation methods based on the sequence-to-sequence model, space weather forecast texts can be automatically generated. To conduct our generation tasks at a fine-grained level, a taxonomy of space weather phenomena based on descriptions is presented. Then, our MDH (Multi-Domain Hybrid) model is proposed for generating space weather summaries in two stages. This model is composed of three sequence-to-sequence-based deep neural network sub-models (one Bidirectional Auto-Regressive Transformers pre-trained model and two Transformer models). Then, to evaluate how well MDH performs, quality evaluation metrics based on two prevalent automatic metrics and our innovative human metric are presented. The comprehensive scores of the three summaries generating tasks on testing datasets are 70.87, 93.50, and 92.69, respectively. The results suggest that MDH can generate space weather summaries with high accuracy and coherence, as well as suitable length, which can assist forecasters in generating high-quality space weather forecast products, despite the data being starved.

Key words Space weather, Deep learning, Data-to-text, Natural language generation

Classified index P162

0 Introduction

Space weather is a consequence of the behaviors of the Sun, the nature of the Earth's magnetic field and atmosphere, and our location in the solar system. The state of space weather changes all the time. Severe space weather phenomena could cause damage to spacecraft, and even cause great economic and social loss. Analyzing the sources of space weather phenomena and forecasting the future of space weather are significant channels for defending against space weather disasters^[1,2].

Space weather forecast includes an artificial forecast and an objective model forecast. A space weather artificial forecast product contains a summary of space weather in the past 24 h and a space weather forecast in 1~3 days. The work in this paper is aimed at the summaries of the artificial forecast. In the summary texts, changes in space weather in the past 24 h are described through the analysis of a large amount of space weather observed data. Nowadays, the forecast business is constantly refining. To generate space weather summaries, forecasters need to refer to a large amount of observed

* Supported by the Key Research Program of the Chinese Academy of Sciences (ZDRE-KT-2021-3)

Received February 22, 2023. Revised April 19, 2023

data, analyze their logical relationship, then generate text descriptions in a limited time, which is a heavy workload and affects the business efficiency. In addition, personal experience bias may lead to human-avoidable errors such as inaccurate descriptions. Under these circumstances, there is a need for an artificial intelligence system that can automatically analyze massive observed data and generate accurate, comprehensive, and readable text that can reflect the logical relationship of various space weather phenomena and describe the state of space weather, assisting forecasters in generating summaries.

In the natural language generation field, data-to-text systems can analyze massive data and automatically generate readable texts. Some systems have been developed for the weather report field^[3–5]. However, due to the large amount and great quantity of dimensions of space weather observed data and limited training samples, some issues occur when using prevalent data-to-text models: descriptions of weather phenomena contain incorrect key information (*e.g.*, an incorrect solar active region number), underfitting (a same error phrase appeared in multiple testing samples) and some significant phenomena may be missing in summaries. Besides, prevalent automatic evaluation metrics concentrate on counting how many *n*-grams or subsequences between generating texts and reference texts coincide, ignoring the semantic and logical information of summaries.

To resolve these issues, a description-based taxonomy of space weather phenomena is presented to conduct generation tasks at a more fine-grained level. Besides, other feature engineering approaches including down-sampling, delexicalization, prompt engineering, and data augmentation are done to reproduce factual details ac-

curately and avoid underfitting. Then the MDH (Multi-Domain Hybrid) model is proposed, which contains several sequence-to-sequence-based sub-models to conduct the generation tasks in two stages (generating by domain and concatenating) so that the input sequence in each inference is shortened and the problem that some critical facts may not be selected is alleviated. In addition to automatic evaluation, a human evaluation study is conducted to comprehensively assess generated space weather summaries, including coverage and accuracy. The results show that although the sizes of our datasets are small, the MDH model can automatically generate space weather forecast products with high coverage, accuracy, and coherence. The description of the experimentation is shown in Figure 1.

1 Related Work

Deep neural networks and sequence-to-sequence models have been explored for data-to-text systems in recent years. Mei *et al.*^[6] improved the vanilla sequence-to-sequence model^[7] and presented a model with an encoder-aligner-decoder framework based on LSTM (Long Short-Term Memory)^[8], improving performances of sequence-to-sequence models on content selection in several data-to-text tasks, but the sequence-aligned model structure makes it unsuitable for multi-dimensional input data. Gu *et al.*^[9] incorporated a copying mechanism in sequence-to-sequence learning, addressing an important problem that certain out-of-vocabulary words in the input sequence could not be reproduced correctly in the output sequence and Sam *et al.*^[10] presented several data-to-text corpora and proved that the copying mecha-

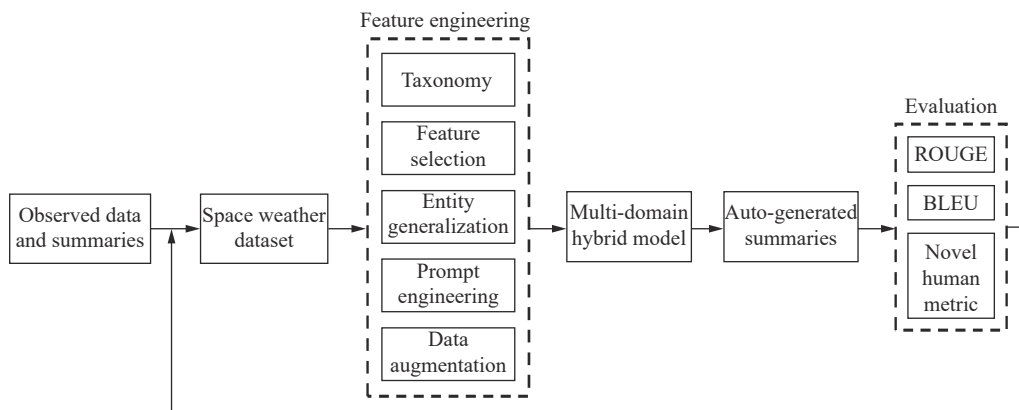


Fig. 1 Description of the experimentation

nism out-performed a series of state-of-the-art neural methods into their data-to-text tasks. However, this copying mechanism only performs well under scenarios with a few entities. Ma *et al.* [11] considered the low-resource data-to-text generation scenario, alleviating the problem of few training samples resulting in underfitting by splitting the generation into two stages: key fact prediction and surface realization, but the structure of this model is not appropriate when input sequence incorporates time sequence. Gong *et al.* [12] improved text correctness under few-shot scenarios with the content matching task, which aligns the structured data and information in the generated text by reconstructing the structure of GPT-2 (Generative Pre-trained Transformer 2) [13]. However, this approach is not appropriate for scenarios where input data has multiple structures. Su *et al.* [14] proposed a Prototype-to-Generate framework, improving data-to-text generation under the few-shot scenario with few training samples through an information retrieval system and a novel prototype selector. Nevertheless, the online retrieval system does little to improve the generation of daily reports.

2 Dataset

RSGA (Report of Solar and Geophysical Activity) products from NOAA (National Oceanic and Atmospheric Administration) are typical daily space weather forecast products, which have a long lifespan, and a wide variety of phenomena mentioned. Every RSGA product contains an analysis of solar active regions and activity, and a geophysical activity summary in the past 24 h. Thus, summary texts are collected from RSGA products instead of Chinese products as our training targets. Both summary texts and space weather observed data taken as input features are available from NOAA*. The collected observed data contains solar active regions, flare

records, GOES proton flux 5-minute data, proton event records, and estimated Planetary K-index 3-hour data. Summary generation for other space weather observed data such as electron flux and solar wind can be implemented through our generalized approach.

3 Feature Engineering

3.1 Description-based Taxonomy

Prevalent sequence-to-sequence-based models have difficulty processing long sequences due to out-of-memory and data redundancy, with the result that these models will generate incorrect key facts or miss some significant information. Thus, for the purpose of conducting generation tasks at a more fine-grained level, a description-based taxonomy is presented based on analyzing the characteristics of the description of various space weather phenomena in the historical review, separating space weather phenomena into three domains: short-term, long-term, and causal relationship (See Table 1).

From each domain, a typical type of space weather phenomenon is selected to conduct the experiment: flares and solar active regions, ≥ 10 MeV proton events, and geomagnetic storms caused by high-speed coronal hole stream, as described in what follows.

3.1.1 Short-term Domain

For phenomena belonging to the short-term domain, forecasters always concentrate on impulsive records or status of the phenomena within 24 h when writing space weather summaries, disregarding the situations in an earlier time (*e.g.*, flares and solar active regions, see Figure 2. Corresponding summary from RSGA product: Solar activity has been at high levels for the past 24 h. The largest solar event of the period was an X2 event observed at 17/2016 Z).

In this domain, the world state contains detailed information about solar activities and the summary texts

Table 1 Taxonomy of space weather phenomena, separating space weather phenomena into three domains

Domain	Typical phenomena
Short-term	Flares and solar active regions, sudden impulses, solar wind, ≥ 2 MeV electron flux
Long-term	≥ 10 MeV proton events, forrush decrease events
Causal relationship	Geomagnetic storms caused by high-speed coronal hole stream

* <https://ftp.swpc.noaa.gov>

are from part analysis of solar active regions and activity of RSGA products. Features used in this domain are shown in Table 2.

To create a challenging dataset, RSGA products over 652 days in 1999–2001 are collected because a large number of solar activities occur in solar activity durations of high periods and the analysis of solar active regions and activity from RSGA is more complex than that in low periods. The corpus is randomly split into 521 samples for training, and 131 samples for testing.

3.1.2 Long-term Domain

For phenomena belonging to the long-term domain, forecasters describe the whole process of the phenomena in space weather summaries. Some characteristics of the phenomena such as beginning time, peak time, end time, and changing trend may be mentioned in the summary, and the phenomena can start within the past 5 days. (*e.g.*, ≥ 10 MeV proton events see Figure 3. Corresponding summary from RSGA product: The greater than 10 MeV

proton event that began at 24/12:15 UT continued to decline today).

In this domain, the world state is represented by a set of observations and events related to the greater than 10 MeV proton. Features used in this domain are shown in Table 3.

The greater than 10 MeV proton event summary texts are from part geophysical activity summary of RSGA products. Since proton events are occasional (even in solar activity durations of high periods), RSGA products over 772 days from 1999 to 2012 were collected (620 samples for training, and 152 samples for testing).

3.1.3 Causal Relationship Domain

For phenomena belonging to the causal relationship domain, forecasters describe the sources of the phenomena (if the sources were found) in space weather summaries. For example, the source of a geomagnetic storm can be analyzed by the composite coronal hole index P_{CH} [15] in the past 72 h. If the source of a geomagnetic storm is a

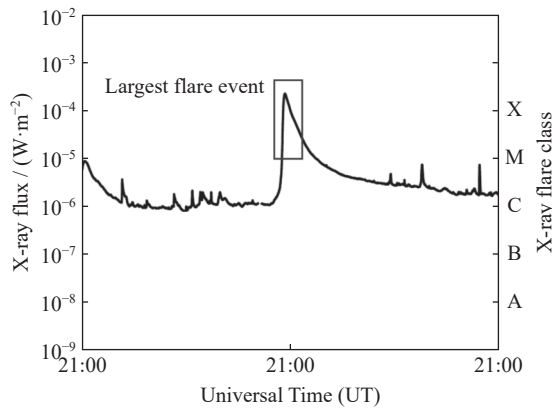


Fig. 2 GOES X-ray flux during 21:00 UT from 16 to 18 February 2023

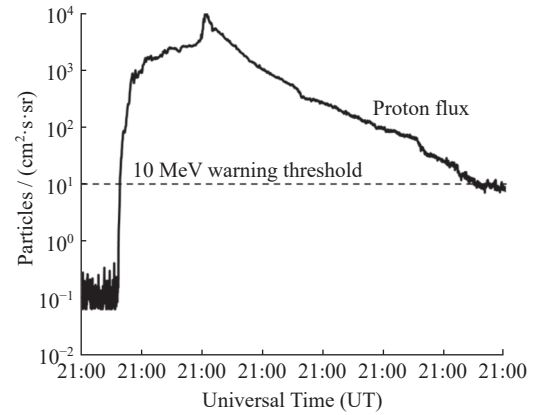


Fig. 3 GOES SEM Proton flux during 21:00 UT from 23 to 30 September 2021

Table 2 Features of flare events and solar active regions

Input features	Details
Solar active regions	Solar active regions in the past 24 h, containing area, evolution, serial number, coordinates
Flares	Flare records in the past 24 h, containing grade, duration, peak time, serial number, and coordinate of the source region

Table 3 Features of greater than 10 MeV proton events

Input features	Details
≥ 10 MeV proton flux	GOES proton flux 5-minute data in the past 24 h
Proton event record	Last proton event record, containing beginning time, peak flux, peak time, end time (if ended)

high-speed coronal hole stream, it will be mentioned in the space weather summary.

In this domain, the world state is represented by Planetary K -index observed data and the source of geomagnetic storm events (See Table 4).

The attribution feature of each sample can be deduced by the composite coronal hole index P_{CH} in the past 72 h. Since the analysis of the relationship between P_{CH} and high-speed coronal hole stream is not the focus of our work, the attribution feature is given with a strong causal relationship. The geomagnetic storm event summary texts are from part geophysical activity summary of RSGA products. To create the corpus, RSGA products over 987 days (2007–2009) are collected, which are randomly split into 769 samples for training, and 218 samples for testing.

3.2 Down-sampling Based Feature Selection

In order to avoid the model input sequence being too long, and ensure that our model can accurately capture the key facts and trend of space weather phenomena (e.g., greater than 10 MeV proton fluxes continued to decline today, fluxes of greater than 10 MeV protons began to increase), the feature selection is done. This part of the work is to extract useful features from the observed data to reduce the complexity and redundancy of input data.

Specifically, down-sampling is conducted on time series data: to reduce the frequency of proton flux sequences. For the purpose of retaining as many event trend characteristics as possible after down-sampling, a novel automatic time sequence down-sampling algorithm is presented (See Figure 4), which takes both extremum points of time sequences and data continuity into consideration. Given a proton flux 5 min data in the past 24 h and a target length, every extremum point in the 5 min sequence is found, and its significance is evaluated. The extremum point list is then truncated using the significance values. The extremum point significance is defined as the distance between the extremum and the mean value multiplied by the mean of the left

and right differences. Then a sample point is inserted in the medium of the longest not-sampled interval until the number of sampled points reaches the target length. In this way, a low-frequency sequence with an accurate trend of the proton event can be obtained.

A univariate experiment is conducted to explore the appropriate target length of the proton flux on our down-sampling algorithm according to the matching rate of trending words between the generated summary and reference. Given a pair of reference and automatic generation, if both reference and generation have descriptors (See Table 5) reflecting the same changing trend pattern of proton events, this pair of texts match. These descriptors are induced by analyzing the artificial reference text.

The result shows that when the target length is set around 18, automatic generations' description of proton events has a relatively high matching rate with corre-

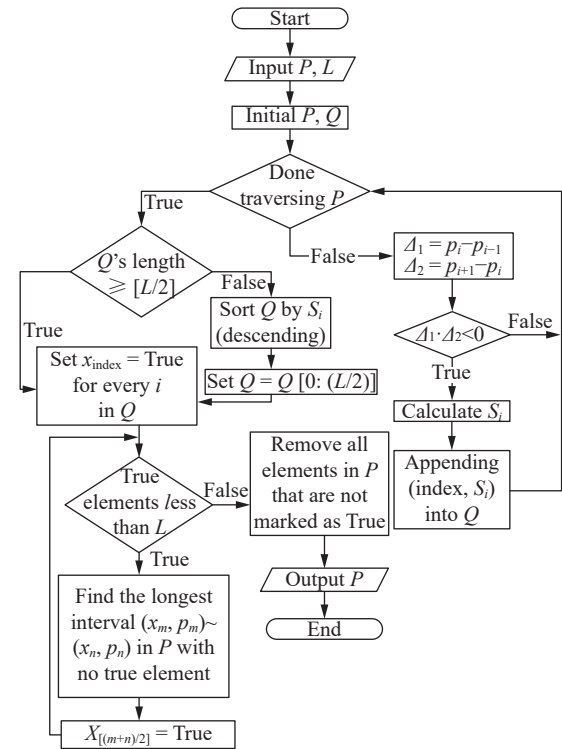


Fig. 4 Flow chart of our automatic time sequence down-sampling algorithm

Table 4 Features of geomagnetic storm events

Input features	Details
Planetary K -index	Estimated planetary K -index 3-hour data in last 24 h
Attribution	An integer value: 1 stands for high-speed coronal hole stream is the source of the event 0 stands for not unrelated to any high-speed coronal hole stream

sponding references (See Figure 5). Therefore, every proton flux sequence is down-sampled to narrow the length from 288 to 18. A sample of this approach is shown in Figure 6, which shows that our algorithm can well preserve the trend of proton events while reducing the frequency.

3.3 Entity Generalization

To address the problem that some key information of phenomena mentioned in summaries is incorrect, delexicalization is conducted to generalize infinitely-valued entities in the corpus. For flare events, and solar active regions, infinitely-valued arguments such as the serial numbers of the active regions, the coordinates of the active regions, and the beginning times of events are delexicalized (using placeholder tokens to replace the values)

Table 5 Descriptors of proton event trend patterns

Pattern	Trend descriptors
Increase	increase, rise, exceed
Decrease	decrease, decline, decay, drop

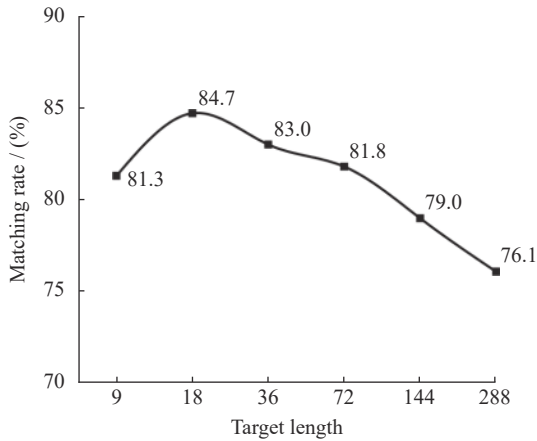


Fig. 5 Result of the univariate experiment for exploring the appropriate target length on down-sampling

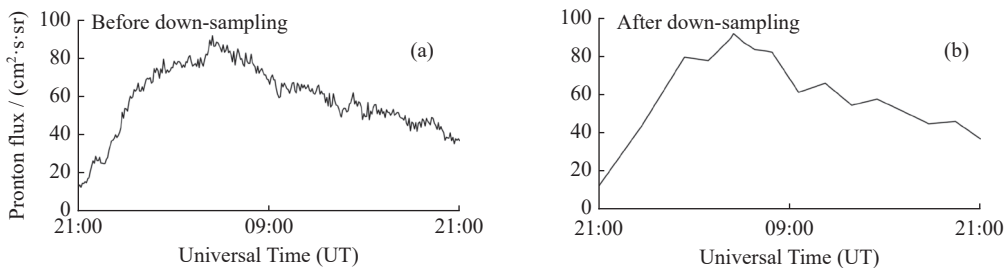


Fig. 6 Proton flux from 21:00 UT 10 to 11 January 2002

in both the world state and training targets. Similarly, for greater than 10 MeV proton events, entities including beginning time, peak flux, peak time, end time (if ended) are extracted, which are delexicalized in the summaries.

The reason why delexicalization can be done is that such delexicalized arguments are verbatim inserted in the summary texts eventually, thus the ultimately generated summary texts are not affected. This method narrows down the size of the vocabulary and prevents the sparsity issue of the embedding layer so that the model can make a more robust performance^[16,17].

3.4 Prompt Engineering

He *et al.*^[18] have proved that, when combined with language models, prompts such as single tokens are effective in guiding the model to plan what to describe while training on corresponding data with specific attributes. Therefore, four kinds of markers: very low, low, moderate, and high are designed to represent solar activity levels, which are decided by the flare records in the past 24 h (See Table 6). These special markers are prepended to the corresponding source input sequences to help our model with selecting flare records and content planning.

3.5 Data Augmentation

To achieve few-shot learning and avoid underfitting, we conduct data augmentation to expand the number of training samples.

Although over 600 samples targeted at the greater than 10 MeV proton are collected for training, the number of the samples reaching proton events is not enough. Therefore, sample augmentation is conducted: to create new training samples, the ≥ 10 MeV proton flux of the original training sample is shifted for s steps, s is a random integer from the range $-4 \sim 4$, then a random ratio from the range $0.845 \sim 1.155$ is given to the time sequence. Through this method, the number of training sam-

ples is expanded to 2233.

Similarly, the number of samples reaching storm events is not enough (especially major storm events and severe storm events). Therefore, sample augmentation is conducted: to create new training samples, the 3 h Planetary K -index sequence is shuffled. By this means, the number of training samples is expanded to 6054.

4 Multi-domain Hybrid Data-to-text Model

For the purpose of avoiding the situation that prevalent sequence-to-sequence models would face an out-of-memory problem and lose some significant information from input sequences when training with long sequences inputs or long text targets and reduce data redundancy, we propose a Multi-Domain Hybrid (MDH) model. In this work, one BART (Bidirectional Auto-Regressive Transformers) pre-train model^[19] is used for generating flares and solar active regions summaries, and two distinct Transformer models^[20] are used for generating greater than 10 MeV proton event summaries and geomagnetic storm event summaries respectively. Figure 7 shows the framework of the MDH model.

Our MDH model contains three sequence-to-sequence-based sub-models (one BART pre-trained model and two Transformer models). This novel model sepa-

rates the generation into two stages: sub-models first generate summary text for each domain respectively, and then generated texts are concatenated in a fixed order (first flares and solar active regions, next geomagnetic storm events, then proton events) as the complete space weather summary texts, because artificial summary products always present a fix narrative order.

4.1 BART Pre-trained Sub-model

Since the fact that size of our dataset is small, and summaries of flares and solar active regions have the characteristics of a large number of descriptions and long length, it is difficult for a randomly initial neural network to learn the grammar and coherence of the summaries. While the BART model has been trained with rich corpus and advanced noisy training methods, getting state-of-the-art performances at several well-known data-to-texts tasks in recent years, thus a BART pre-trained model is fine-tuned to generate summaries of flares and solar active regions. Containing a bidirectional encoder and an autoregressive decoder, BART is a sequence-to-sequence-based model. The input of the model is the linearized representation of the space weather observed data, and the output is the summary texts. Figure 8 shows the structure of the BART model.

Compared with other pre-train models for generation tasks, such as T5^[21], BERT for Sequence Generation^[22], several advanced noisy methods including text

Table 6 Decision rules of solar activity level in the past 24 h

Flare	Activity level
No flare is higher than B grade	Very low
Highest flare is at C grade	Low
Number of flares at M1.0~M4.9 is between 1~4, and there is no flare higher than M4.9	Moderate
Number of flares at M1.0~M4.9 is over 4, or there is a flare over M4.9	High

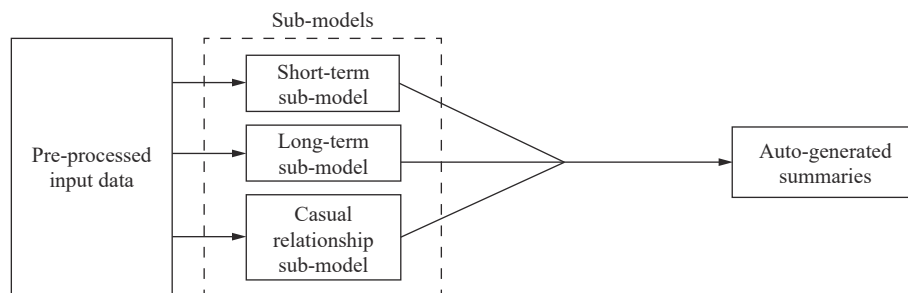


Fig. 7 Framework of MDH model

infilling, and document rotation are used to corrupt the training texts during the pre-training. In our work, a base-size BART pre-trained model is used: 768-dimensional word embeddings are used in word embedding layers, and 6 transformer layers are used in both the encoder and the decoder. In each transformer layer, linear layers with 768-dimensional hidden states are used. The output linear layer also has 768-dimensional hidden states.

4.2 Transformer Sub-model

With the self-attention mechanism, the Transformer model has been proven to have excellent long-distance feature capture capability and parallel computing capability^[20]. Considering the fact that both summaries of greater than 10 MeV proton events and geomagnetic storm events have simpler narrative logic than that of flares and solar active regions, and their need to analyze numeric data, two distinct Transformer models are used for proton events summaries and geomagnetic storm events summaries generation. Figure 9 shows the structure of the Transformer model.

Both two Transformer sub-models we use in the

different domains have the same architecture: the input linear layer has 128-dimensional hidden states, 128-dimensional word embeddings are used as the output embedding layer, and 2 transformer layers are used in both the encoder and the decoder. In each transformer layer, output linear layers with 512-dimensional hidden states are used.

5 Experimental Setup

The experiment is conducted using our novel dataset and MDH model. Several details are highlighted as follows.

Backpropagation is taken as the learning method^[23], Adam^[24] as the optimizer, and a batch size of 8 is used for training. The loss L for each timestep t is the negative log likelihood of the ground-truth x'_t for that timestep:

$$L_t = -\lg P(x'_t).$$

Together with the loss of all timesteps T , the loss of the model becomes

$$L = \sum_{t=1}^T L_t.$$

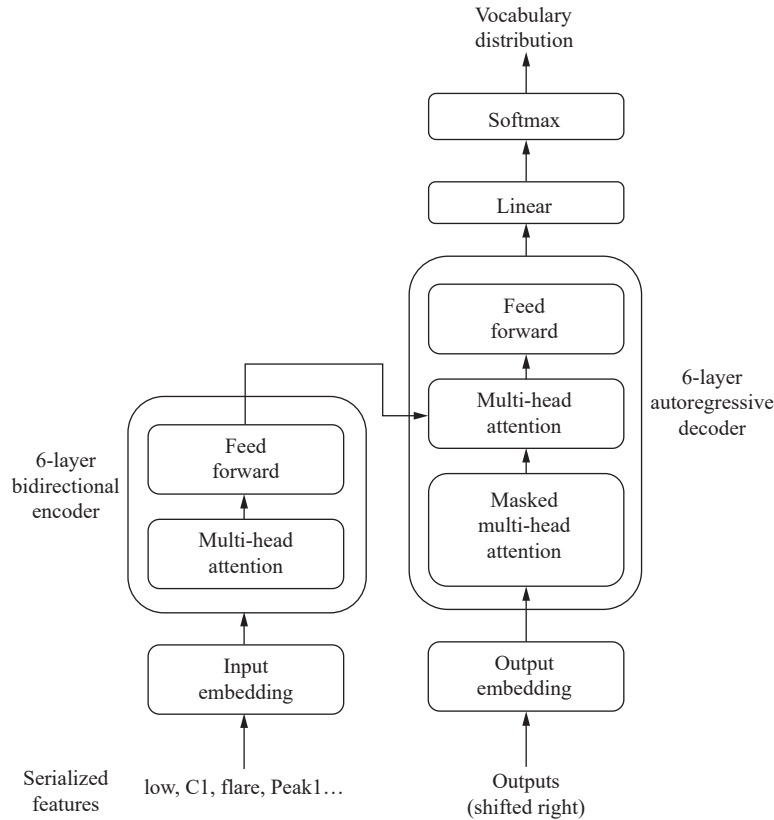


Fig. 8 Structure of the BART model (Serialized inputs consist of solar activity level, flare events and active regions)

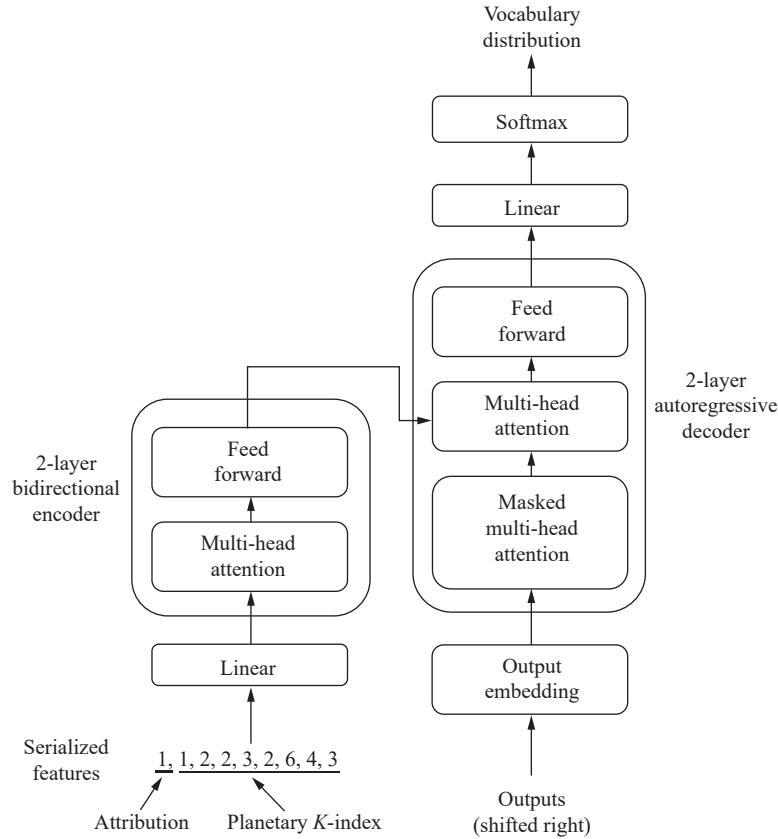


Fig. 9 Structure of the Transformer model (This model takes attribution and 3-hour planetary K -index as inputs for generating geomagnetic storm event summary)

As for the decoding strategy during inference, greedy search is used to generate summaries. Greedy search is a classic strategy that can generate steady summaries. At each timestep, choose the most possible token as the input of the next timestep from output probability. Compared with other prevalent decoding strategies such as top- k sampling and beam search, greedy search has lower space complexity^[25], and thus can generate summary texts in a fast way with steady content.

6 Evaluation

Evaluation metrics used for natural language generation tasks have two domains: auto metrics and human metrics. Although automatic evaluation metrics cost humans little effort and the results of them are repeatable, these metrics could not identify how many space weather phenomena mentioned in reference were also mentioned in model generations and how many were contradictory. Besides, automatic evaluation metrics could not accurately evaluate the correctness of the model genera-

tions. It has been proved that human evaluations gave the best insight into the performance of our models^[26]. Thus, several classic automatic evaluation metrics are used and a novel human evaluation metric is presented to examine how well our models performed in terms of the space weather observed data and reference texts.

6.1 Automatic Evaluation Metrics

We use ROUGE (Recall-Oriented Understudy for Gisting Evaluation)^[27] and BLEU (Bilingual Evaluation Understudy)^[28] to automatically evaluate the surface realization of our models. Both ROUGE and BLEU are prevalent evaluation metrics for data-to-text tasks^[26]. They are based on counting the number of overlapping n -grams or subsequences between a reference sequence and a candidate sequence.

6.1.1 ROUGE

ROUGE measures calculate the n -gram recall and subsequence recall between the reference and model generation. Two different ROUGE measures are used: ROUGE-1, ROUGE-L. The calculation is as follows:

$$R_1 = \frac{\sum_{r \in R} \sum_{u \in r} C_1(u)}{\sum_{r \in R} \sum_{u \in r} C_2(u)},$$

where R_1 stands for ROUGE-1 score, r stands for a reference, R stands for a set of reference summaries, u stands for a unigram in references, $C_1(u)$ stands for the maximum of unigrams co-occurring between model generation and reference, $C_2(u)$ stands for the frequency of a unigram occurring in the reference.

$$R = Q/m,$$

$$P = Q/n,$$

$$R_L = \frac{2RP}{R+P}.$$

Here R_L stands for ROUGE-L score, R stands for recall, P stands for precision, m is the length of model generation, n is the length of reference summary and Q stands for the longest common subsequence of reference summary and model generation.

6.1.2 BLEU

BLEU calculates the unigram precision between the reference and model generation. The calculation of BLEU is as follows:

$$p = \frac{\sum_{g \in G} \sum_{u \in g} C_3(u)}{\sum_{g \in G} \sum_{u \in g} C_4(u)};$$

$$P = \begin{cases} 1, & m \geq n, \\ e^{1-n/m}, & m < n; \end{cases}$$

$$B = Pp.$$

Here B stands for BLEU score, P is the brevity penalty, p stands for the unigram precision score, g stands for a model generation, G stands for a set of model generations, u stands for a unigram in model generations, $C_3(u)$ is the minimum between the frequency of a unigram occurring in the model generation and the maximum frequency of a unigram occurring in the reference summary, $C_4(u)$ is the frequency of a unigram occurring in the model generation, m is the length of the model generation and n is the length of reference summary.

6.2 Human Evaluation Metric

To calculate the human evaluation scores, three evaluated description features are presented: accuracy, coverage, and sentence length, as described in what follows.

6.2.1 Accuracy

Accuracy refers to the percentage of the phenomena correctly described in the model generation in terms of the space weather observed data. The calculation of accuracy is as follows:

$$A = \frac{T}{W_g},$$

where A stands for accuracy score, W_g stands for the number of phenomena described in model generation, T stands for the number of true phenomena described in model generation.

6.2.2 Coverage

Coverage refers to the extent of reference covered by the model generation. This feature is similar to the work presented by Cascallar-Fuentes *et al.* [29]. The calculation of coverage is as follows:

$$C = \frac{I}{W_r},$$

where C stands for coverage score, W_r stands for the number of phenomena described in the reference, and I stands for the number of phenomena described in both reference and model generation. Therefore, a full coverage score (100%) will be obtained when all the phenomena mentioned in the reference are described by the model generation.

6.2.3 Sentence Length

An evaluation metric should require that an automatically generated sentence not be too short [28]. Sentence length refers to the comparison of the length of sentences between reference and model generation, preferring longer sentences. The calculation of sentence length is as follows:

$$S = \max \left\{ \frac{L_g}{L_r}, 1 \right\},$$

where S stands for sentence length score, L_r stands for the length of the sentences from reference describing the same phenomena in reference and model generation, L_g stands for the length of the sentences from model generation describing the same phenomena in reference and model generation.

6.2.4 Score

Having these three values: accuracy, coverage, and sentence length, calculate a score based on them:

$$H = 0.4A + 0.4C + 0.2S,$$

where H stands for human evaluation score, A stands for accuracy score, C stands for coverage score, S stands for sentence length score. Since the sentence length feature only concentrates on the partially generated sentences, a smaller weight is assigned to this value than others. For each task, 30 distinct records are randomly selected from the testing dataset and asked raters to give the scores. Each selected record was rated 3 times by distinct raters. This setup is similar to former work^[30–31].

6.3 Comprehensive Metric

Since there is an inconsistency between automatic metrics and the human metric, a comprehensive metric is presented by taking the weighted sum of the automatic evaluation scores and the human evaluation score. Due to the fact that automatic metrics are uninterpretable^[32] and human evaluations are typically viewed as the most important form of evaluation for natural language generation systems^[26], a bigger weight is assigned to the human metric than automatic metrics.

$$C = 0.25 \times \frac{R_1 + R_L + B}{3} + 0.75H,$$

where C stands for a comprehensive score, R_1 stands for ROUGE-1 score, R_L stands for ROUGE-L score, B stands for BLEU score, and H is the human evaluation score.

7 Results and Analysis

7.1 Automatic Evaluation Results

The results of the automatic evaluation are shown in Table 7, which shows that our models get outstanding automatic evaluation scores in generating ≥ 10 MeV proton event summaries and geomagnetic storm event summaries, while these three automatic evaluation metrics can't illustrate that our model performs well in generating flares and solar active regions summaries. Considering the fact that forecasters responsible for the RSGA products have different selection preferences among various minor flares (lower than M grade) and solar active regions due to experience bias, our model can't learn the selection preference with high consistency because of the biased training targets. To analyze the effect of experience bias, the ablation of minor solar activities is presented. The result of the ablation is shown in Table 8,

which shows that automatic evaluation scores are greatly improved with the ablation of minor solar activities. In addition to experience bias, the different writing styles of the training samples also have a big impact on automatic evaluation. RSGA products are written by different forecasters in rotation, and the writing styles of different forecasters have distinctive characteristics, especially for solar activities that produce multiple records of flares and active regions every day. A language model that has learned multiple writing styles has difficulty generating text that is consistent with the style of the corresponding reference, causing the automatically generated summaries to fail to achieve high scores on automatic metrics that mainly examine the overlap of n -grams or subsequences. For example, although a single C-class event occurred, a C1 from region 9350 (N18 E30), at 0618 Z. and Region 9350 (N18 E30) produced a C1 flare at 0618 Z. describe the same content, the inconsistency of the writing style results in only a few overlapping of n -grams or subsequences between this pair of texts (generated by the forecaster and MDH model respectively).

In contrast to the task of generating summaries of flares and solar active regions, summaries of proton events have the best automatic evaluation scores, because proton events are sporadic events: when no proton event occurred in the past 24 h, forecasters often do not describe the ≥ 10 MeV proton fluxes (a single word “none” is used to represent the content of the summary in this circumstance), which leads to lots of testing samples without proton events achieve a full score in all au-

Table 7 Results of automatic evaluation

Task	ROUGE-1	ROUGE-L	BLEU
Flares and solar active regions	38.69	36.60	33.01
≥ 10 MeV proton events	96.32	96.20	96.13
Geomagnetic storm events	79.63	83.73	76.93

Table 8 Result of minor solar activities' ablation

Minor solar activities	ROUGE-1	ROUGE-L	BLEU
With	38.69	36.60	33.01
Without	64.11	61.29	61.61

tomatic metrics (both reference and auto-generation are represented as none). For the testing samples with presences of ≥ 10 MeV proton events, further analysis is conducted in the Human Evaluation Results section.

As for the summaries of geomagnetic storm events, forecasters will describe the geomagnetic conditions regardless of whether there have been geomagnetic storm events in the past 24 h. Differences in writing style prevent summaries of geomagnetic storm events from scoring as high in automatic metrics as proton events (*e.g.*, Geomagnetic field activity was at quiet to unsettled levels. The geomagnetic field was quiet to unsettled).

7.2 Human Evaluation Results

The results of the human evaluation are shown in Table 9 and Figure 10, which show that our model generations of flares and active regions summaries have high accuracy, but prefer to describe fewer phenomena than references (*e.g.*, some minor flare events may be missed, as is common in actual forecasting work). In generating ≥ 10 MeV proton events summaries and geomagnetic storm events summaries, our model performs well in terms of all three human description features thus the scores are both very high, which is consistent with the results of the automatic evaluation.

The confusion matrix of ≥ 10 MeV proton events summaries is shown in Table 10 (Positive indicates the presence of a proton event in the past 24 h, Negative indicates the absence), which shows that auto-generated summaries can provide high-precision descriptions of proton events that occurred in the past 24 h. However, Figure 10 shows that a few testing samples describe inaccurate event details of short proton events which last only a few hours due to the low proportion of the samples describing short proton events. In the future, the training samples will be further equalized with sampling techniques to improve the performance on short proton events. In addition, four testing samples have low coverage values because auto-generated summaries miss the description of proton fluxes. Although none of the four

samples reach ≥ 10 MeV proton event level, forecasters still describe the proton fluxes in summaries (*e.g.*, the

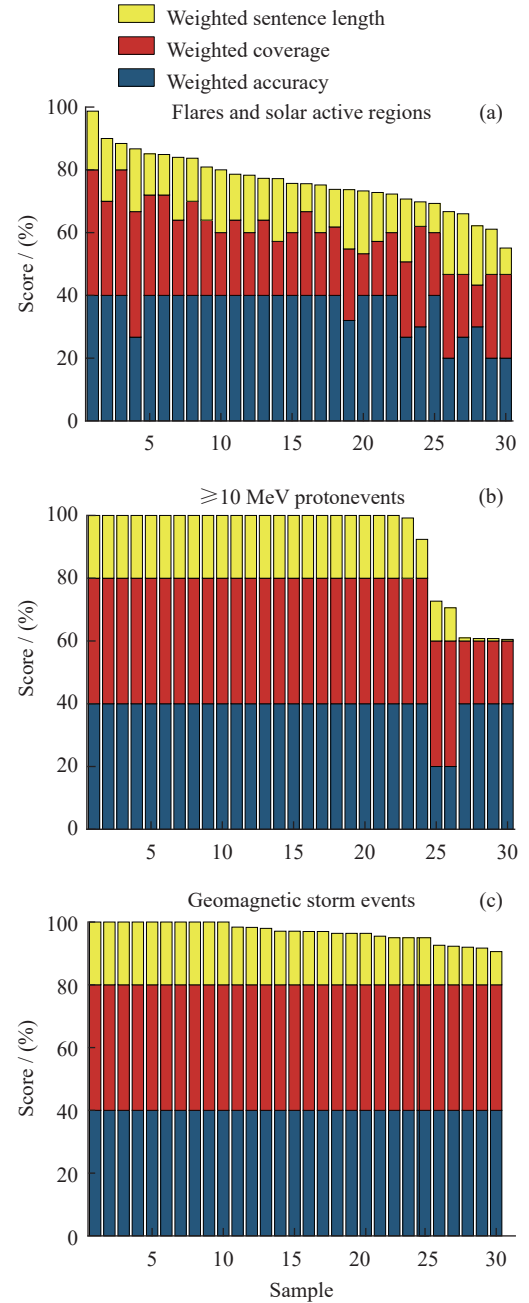


Fig. 10 Distribution of human evaluation scores. Each stacked column represents the human score of a test sample

Table 9 Results of human evaluation

Task	Accuracy	Coverage	Sentence length	Score
Flares and solar active regions	89.34	77.95	77.75	82.47
≥ 10 MeV proton events	96.67	93.33	82.97	92.59
Geomagnetic storm events	100.00	100.00	84.46	96.89

Table 10 Confusion matrix of ≥ 10 MeV proton events summaries

	Observed positive	Observed negative
Described positive	8	0
Described negative	0	22

greater than 10 MeV proton flux at geosynchronous orbit stayed below 10 pfu all period), which occasionally happens in actual forecast business. Therefore, this kind of omission can be tolerated.

Summaries of geomagnetic storm events have the best human evaluation scores because the attribute features in our dataset are given with strong causalities, guiding MDH model to generate summaries with strong completeness and accuracy. The accuracy value and coverage value in generating geomagnetic storm events summaries may be hard to get full scores in practical applications.

Coverage values are generally not higher than accuracy values among these three tasks because the training samples are written by different forecasters, and the selection preference shown by our trained model is generally the intersection of the individualized selection preferences of different forecasters. Thus, summaries generated by our model may describe less content than the corresponding references.

7.3 Comprehensive Scores

The results of comprehensive scores are shown in Table 11, which shows that MDH model performs well on those three tasks, especially in generating summaries of ≥ 10 MeV proton events and geomagnetic storm events.

7.4 Case Study

In this section, three pairs of comparisons of model generation and reference in each domain are presented as follows.

Example 1

Reference: solar activity was moderate due to a single isolated M1 event from region 8791 (S14 E53) at 07/21:20 UT. Only infrequent small C-class flares occurred during the rest of the day. New region 8792 (N17 E69) was numbered during this period.

Auto-generation: Solar activity was moderate. The largest event was an optically uncorrelated M1 flare at 2120 Z. New region 8792 (N17 E69) was numbered today.

Table 11 Comprehensive scores

Task	Comprehensive score
Flares and solar active regions	70.87
≥ 10 MeV proton events	93.50
Geomagnetic storm events	92.69

Example 1 shows that our model can describe the solar activity level (moderate), largest event (M1), and new region (region 8792) accurately while missing some minor C-class flare events and the source of the largest flare described in the reference, which is acceptable for actual artificial forecasts.

Example 2

Reference: the greater than 10 MeV proton event that began at 24/12:15 UT continued to decline today.

Auto-generation: The greater than 10 MeV proton event continues in progress. The present particle flux is about 269.0 pfu and is slowly declining.

Example 2 shows that our model can precisely describe the changing trend of the greater than 10 MeV proton event (declining). Although a detail is missing (began at 24/12:15 UT), an additional one is added (the present particle flux is about 269.0 pfu).

Example 3

Reference: The geomagnetic field was predominantly quiet to unsettled. A single period of minor storming occurred due to a coronal hole high speed stream.

Auto-generation: The geomagnetic field ranged from quiet to minor storm levels due to the influence of a recurrent coronal hole high speed stream.

Example 3 shows that our model can describe the geomagnetic storm accurately (minor storm) and explain its source of it (coronal hole high speed stream).

8 Discussion

As the combination of artificial intelligence and linguistic methods, natural language generation methods are used to analyze observed data and generate summary text automatically in the space weather domain in our work. In this study, the space weather summary generation tasks are conducted at a fine-grained level through the description-based taxonomy and proposed MDH model, resolving several problems in artificial space weather summary automatic generation. The results

based on our novel human metric show that MDH model can automatically select space weather phenomena concerned by forecasters from the space weather observed data and present them on summary texts with high accuracy. Besides, our model generations can basically describe the same phenomena as references with appropriate sentence lengths. Although flares and solar active regions summaries have lower automatic evaluation scores and coverage (77.95%), this is mainly caused by experience bias and different writing styles, which are tolerable in the practical forecasting business. Limited by the experimental environment, the base-size BART pre-trained model is used instead of the largest version. As for the structure of the Transformer sub-model, the number of layers and hidden units is lightly tuned according to the automatic evaluation metrics (ROUGE and BLEU) and the results of pretests suggest that the Transformer model with 2-layer encoder and decoder, 128 hidden units in embedding layers and 512 hidden units in linear layers is suitable for our study.

9 Conclusion

In this paper, the challenges facing artificial space weather summary automatic generation are explored by introducing a new dataset and proposing MDH model and a novel human metric for evaluating generated contents' accuracy, coverage, and sentence length.

In the future, the auto generations will be converted into Chinese products through machine translation approaches. Moreover, experimental prediction of space weather will be conducted to predict future space weather phenomena including solar flares, coronal mass ejections, and particle events by building models based on the sequence-to-sequence framework, then training them with the observed data of space weather, and the solar activity forecast and geophysical activity forecast from RSGA products as new training targets.

Acknowledgments The authors acknowledge the use of RSGA products and space weather observed data from NOAA (<https://ftp.swpc.noaa.gov/pub/>).

References

- [1] LILENSTEN J, BELEHAKI A. Developing the scientific basis for monitoring, modelling and predicting space weather[J]. *Acta Geophysica*, 2009, **57**(1): 1-14
- [2] SINGH A K, BHARGAWA A, SIINGH D, *et al.* Physics of space weather phenomena: a review[J]. *Geosciences*, 2021, **11**(7): 286
- [3] GOLDBERG E, DRIEDGER N, KITTREDGE R I. Using natural-language processing to produce weather forecasts[J]. *IEEE Expert*, 1994, **9**(2): 45-53
- [4] REITER E, SRIPADA S G, HUNTER J, *et al.* Choosing words in computer-generated weather forecasts[J]. *Artificial Intelligence*, 2005, **167**(1/2): 137-169
- [5] XING X Y, WAN X J. Structure-aware pre-training for table-to-text generation[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online: Association for Computational Linguistics, 2021: 2273-2278
- [6] MEI H Y, BANSAL M, WALTER M R. What to talk about and how? Selective generation using LSTMs with coarse-to-fine alignment[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: NAACL, 2016: 720-730
- [7] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]//Proceedings of the 3rd International Conference on Learning Representations. San Diego: Computational and Biological Learning Society, 2015
- [8] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. *Neural Computation*, 1997, **9**(8): 1735-1780
- [9] GU J T, LU Z D, LI H, *et al.* Incorporating copying mechanism in sequence-to-sequence learning[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin: Association for Computational Linguistics, 2016: 1631-1640
- [10] WISEMAN S, SHIEBER S M, RUSH A M. Challenges in data-to-document generation[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017: 2253-2263
- [11] MA S M, YANG P C, LIU T Y, *et al.* Key fact as pivot: a two-stage model for low resource table-to-text generation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 2047-2057
- [12] GONG H, SUN Y W, FENG X C, *et al.* TableGPT: few-shot table-to-text generation with table structure reconstruction and content matching[C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: International Committee on Computational Linguistics, 2020: 1978-1988
- [13] RADFORD A, WU J, CHILD R, *et al.* Language models are unsupervised multitask learners[OL]. [2019-02-15]. http://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- [14] SU Y X, MENG Z Q, BAKER S, *et al.* Few-shot table-to-text generation with prototype memory[C]//Findings of the Associ-

- ation for Computational Linguistics: EMNLP 2021. Punta Cana: Association for Computational Linguistics, 2021: 910-917
- [15] BU X, LUO B, SHEN C, *et al.* Forecasting high-speed solar wind streams based on solar extreme ultraviolet images[J]. *Space Weather*, 2019, **17**(7): 1040-1058
- [16] HOSSEINI-ASL E, MCCANN B, WU C S, *et al.* A simple language model for task-oriented dialogue[C]//Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020: 20179-20191
- [17] ARUN A, BATRA S, BHARDWAJ V, *et al.* Best practices for data-efficient modeling in NLG: how to train production-ready neural models with less data[C]//Proceedings of the 28th International Conference on Computational Linguistics: Industry Track. Online: International Committee on Computational Linguistics, 2020: 64-77
- [18] HE J X, KRYŚCIŃSKI W, MCCANN B, *et al.* CTRLsum: towards generic controllable text summarization[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi: Association for Computational Linguistics, 2022: 5879-5915
- [19] LEWIS M, LIU Y H, GOYAL N, *et al.* BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 7871-7880
- [20] VASWANI A, SHAZEER N, PARMAR N, *et al.* Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017: 6000-6010
- [21] RAFFEL C, SHAZEER N, ROBERTS A, *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer[J]. *The Journal of Machine Learning Research*, 2020, **21**(1): 140
- [22] ROTHE S, NARAYAN S, SEVERYN A. Leveraging pre-trained checkpoints for sequence generation tasks[J]. *Transactions of the Association for Computational Linguistics*, 2020, **8**: 264-280
- [23] RUMELHART D E, DURBIN R, GOLDEN R, *et al.* Backpropagation: the basic theory[M]//CHAUVIN Y, RUMELHART D E. Backpropagation: Theory, Architectures, and Applications. Hillsdale: Lawrence Erlbaum Associates, 1995: 1-34
- [24] KINGMA D P, BA J. Adam: a method for stochastic optimization[C]//3rd International Conference on Learning Representations. San Diego: Computational and Biological Learning Society, 2015
- [25] HOLTZMAN A, BUYS J, DU L, *et al.* The curious case of neural text degeneration[C]//8th International Conference on Learning Representations. Addis Ababa: OpenReview. net, 2020
- [26] CELIKYILMAZ A, CLARK E, GAO J F. Evaluation of text generation: a survey[OL]. arXiv preprint arXiv: 2006.14799, 2020
- [27] LIN C Y. ROUGE: a package for automatic evaluation of summaries[C]//Text Summarization Branches Out. Barcelona: Association for Computational Linguistics, 2004: 74-81
- [28] PAPINENI K, ROUKOS S, WARD T, *et al.* BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia: ACM, 2002: 311-318
- [29] CASCALLAR-FUENTES A, RAMOS-SOTO A, BUGARÍN A. Meta-heuristics for generation of linguistic descriptions of weather data: experimental comparison of two approaches[J]. *Fuzzy Sets and Systems*, 2022, **443**: 173-202
- [30] CHEN Z Y, EAVANI H, CHEN W H, *et al.* Few-shot NLG with pre-trained language model[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 183-190
- [31] LI L, MA C, YUE Y L, *et al.* Improving encoder by auxiliary supervision tasks for table-to-text generation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Bangkok: Association for Computational Linguistics, 2021: 5979-5989
- [32] VAN DER LEE C, GATT A, VAN MILTENBURG E, *et al.* Best practices for the human evaluation of automatically generated text[C]//Proceedings of the 12th International Conference on Natural Language Generation. Tokyo: Association for Computational Linguistics, 2019: 355-368

作者简介



罗冠霆 男, 1998年3月出生于广东省东莞市. 中国科学院大学国家空间科学中心2023届计算机技术专业硕士研究生, 主要研究方向为空间环境预报、自然语言处理.

E-mail: guanting.luo@ist.osaka-u.ac.jp