A Timestep aware Sentence Embedding and Acme Coverage for Brief but Informative Title Generation

Quanbin Wang[1,2][*],Xiexiong Lin[3], Feng Wang[3]
[1]Alibaba Group
[2]College of Computer Science and Technology, Zhejiang University, China
[3]Ant Group
{quanbin.wqb,xiexiong.lxx,zifan.wf}@antgroup.com

Abstract

The title generation task that summarizes article content in recapitulatory words relies heavily on utilizing the corresponding key context. To generate a title with appropriate information in the content and avoid repetition, we propose a title generation framework with two complementary components in this paper. First, we propose a Timestep aware Sentence Embedding (TSE) mechanism, which updates the sentences' representations by re-locating the critical words in the corresponding sentence for each decoding step. Then, we present an Acme Coverage (AC) mechanism to solve the repetition problem and preserve the remaining valuable keywords after each decoding step according to the final vocabulary distribution. We conduct comprehensive experiments on various title generation tasks with different backbones, the evaluation scores of ROUGE and METEOR in varying degrees are significantly outperforming most of the existing state-of-the-art approaches. The experimental results demonstrate the effectiveness and generality of our novel generation framework TSE-AC.

1 Introduction

On account of the existence of many articles and news, automated title generation which summarizes the source content into a succinct title with recapitulatory words can significantly reduce the cost of obtaining information and improve the efficiency of information transmission.

As the example shown in Table 1, the subject is composed of different critical words from the email's first two sentences. When generating the word "birthday", the model needs to focus on the phrase "a year old" to form the first sentence's representation. But for the second target word "party", the model has to update the first sentence's embedding by re-locating to the word "party". Therefore, it is necessary for a title generation method to refresh the sentences' embeddings with corresponding key words in different decoding timestep.

email content	We are planning a **party** it has been a while since the group has had a party and my daughter is going to be **a year old**. So we are planning a **party** for November 14th and Vandhana and I would like to **invite** everyone in research and their family. As yet we do not have the ...
email subject	birthday party invitation

Table 1: An example of email and its subject.

End-to-end neural generation models have achieved impressive performance in title generation via sequence-to-sequence (seq2seq) framework (Bahdanau et al., 2014; Sutskever et al., 2014). However, most of them suffered from the difficulty of integrating key information from different parts of the source content effectively. Some other works have realized the importance of the key components from different pieces of the article content (Gehrmann et al., 2018; Li et al., 2018, 2020a; Tan et al., 2017; Cohan et al., 2018), nevertheless, they extract all potentially useful parts as another static input of the model instead of dynamically locating the critical words in combination with the decoding states to update the sentence level representations.

To overcome the shortness of previous works that freeze the sentences' embeddings without considering the changes of the decoding states. we propose a novel title generation architecture with a TSE mechanism, which finely adopts different related words of each sentence to update the sentence's embedding vector in each decoding step. Furthermore, the architecture with TSE incorporates the commonly used hierarchical encoder (Yang et al., 2016; Tan et al., 2017; Cohan et al.,

[*] Corresponding author.

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1906 - 1918
July 10-15, 2022 ©2022 Association for Computational Linguistics

2018), part of speech (POS) information (Liu et al., 2019), and graph structure (Zhang et al., 2018; Yu et al., 2020) to take full advantages of inner relations among words within each sentence as well as among sentences through the article content, so as to re-locate the related key words for each target word more accurately. This encoding method gives insight into the correlation of each keyword and then integrates them well to summarize the article. To the best of our knowledge, this is the first work which finely updates the embedding of each sentence via the target words for each decoding step.

Besides, repetition is a common problem in generation tasks, which especially stands out in the title generation scenerio due to its succinctness. The coverage mechanism is widely used to address this problem and achieves impressive results (Tu et al., 2016; See et al., 2017). Previous coverage mechanisms focus on maintaining a coverage vector which is the sum of attention distributions over all previous decoding timesteps and make the next attention weights as different from the coverage vector as possible. However, in title generation, the penalization with this kind of coverage vector will penalize all attended keywords, disturbing the attention mechanism in the following decoding steps. And the generated titles may lose some keywords because of the inappropriate penalization. Moreover, since our TSE relies heavily on the attention mechanism, the superimposed negative influence will seriously impair the accuracy of the generated title. In this paper, we leverage an AC mechanism that only prevents repeatedly attending to the word which has been generated actually to avoid producing repetitive text but without missing other valuable keywords.

Our contributions are summarized as follows:

- We propose an end-to-end title generation framework with timestep aware sentence embedding, which is effective for the model to dynamically encode each sentence with critical words and the latest valuable information in the corresponding decoding timestep.

- We present an acme coverage mechanism that solves the repetition problem but avoids unreasonable penalization, which obtains significant outperformance on our novel architecture and other seq2seq models.

- Our model achieves significant improvements over several strong baselines on email subject generation and news headline generation tasks. The detailed experimental results demonstrate that our method is effective and general for different kinds of title generation scenarios.

2 Related Works

Title generation has been investigated for a long time, some classical works (Kennedy and Hauptmann, 2000; Jin and Hauptmann, 2001, 2002; Jin et al., 2020) presented various approaches for different kinds of title generation tasks. Rush et al. (2015) first adopted the attention mechanism to the abstractive title generation task, which is a commonly used method for diverse text generation scenarios (Paulus et al., 2018a; Song et al., 2019; Bi et al., 2020a; Lewis et al., 2020). Zhang et al. (2020) pre-trained a model with objectives tailored for abstractive text summarization, they achieved excellent performances on various tasks. However, it remains a major challenge for seq2seq models to tackle document inputs since more information in long documents will probably confuse the model and result in degraded performance. Some previous works (Zhang and Tetreault, 2019; Tan et al., 2017) chose the approach with two separate stages to avoid long inputs, they generated the title via some selected sentences. These kinds of methods may lose necessary information in the generated titles once some useful sentences are ignored.

Some other works applied hierarchical frameworks to encode document level inputs. Tan et al. (2017); Cohan et al. (2018) utilized a hierarchical encoder to model the discourse structure of long documents, but the sentences' embeddings are static during all decoding steps in their works, we update the embedding of each sentence for every targeted word dynamically based on the corresponding valuable words and the decoding state. Li et al. (2020b) presented a hierarchical model to generate summarization for multiple documents. The main differences between their work and ours are twofold. First is the granularity of the encoder, we pay more emphasize on tokens instead of paragraphs since every word is crucial for title generation. The second but critical difference is the related information locating mechanism used in decoder. They first extracted a central location for the key information and select several paragraphs around it. Our model proposes a TSE mechanism which will focus on all related keywords in each sentence to update their embeddings dynamically

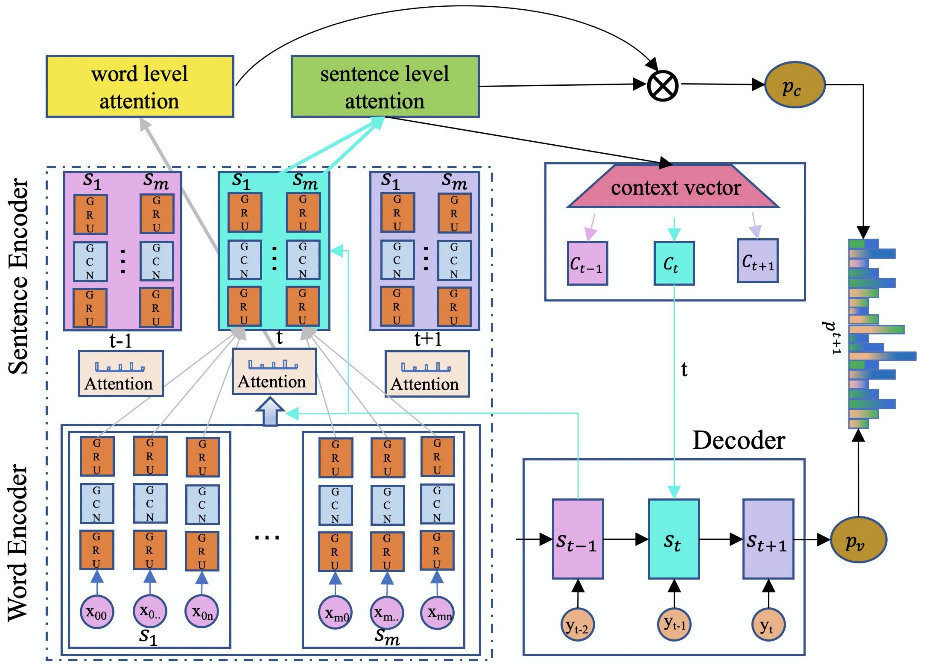

Figure 1: The model architecture of our title generation method

for every decoding step.

Besides, repetition is a common problem in generation tasks, there are several efforts made to address this issue. Temporal attention is a technique that has been applied to neural machine translation (NMT) (Sankaran et al., 2016) and summarization (Nallapati et al., 2016), where each attention distribution is divided by the sum of the previous, which effectively avoids repeated attention but distorts the signal from the attention mechanism and then reduces the performance. N-gram blocking (Paulus et al., 2018b) is another technique proposed to discard the n-gram if it appeared already. Coverage mechanism is a widely used technique to dampen repetition, which was first applied to NMT by maintaining a coverage vector (Tu et al., 2016; Mi et al., 2016). A simple approach that upholds the coverage vector with the sum of the attention distributions was introduced in summarization task and achieved impressive results (See et al., 2017). While this approach penalizes the previous attention distribution indiscriminately, which disturbed the generation of later target words and resulted in degraded performance. We leverage an AC mechanism that maintains a coverage vector based on the attention weights with the highest probability of the final vocabulary, and outperforms difference evaluation metrics compared to the vanilla coverage.

3 Method

3.1 Model Architecture

The problem we focused on in this paper is giving an article A as the content of email, news, story etc., which consists of $|A|$ sentences: $A = [S_1, S_2, ..., S_{|A|}]$, we aim to generate a succinct but informative title of the source content, denoted as a token sequence $Y = [y_1, y_2, ..., y_n]$. The architecture of our approach (as shown in Figure1) is comprised of two essential parts: 1) Timestep aware Sentence Embedding; 2) Acme Coverage.

3.2 Timestep aware Sentence Embedding

The TSE mechanism (Figure 1) is motivated by the fact that each word in the title reflects different parts of the source content (as shown in Table 1). We use the hidden state of the decoder to update the representations of all sentences by re-locating the valuable words for each decoding step. Besides, the roles of words, relations within the same sentence and across different sentences are meaningful for the TSE to identify really useful words. We encode the source article with the POS information, dependency parsing information within each sentence and the lexical relations among all sentences explicitly. The details of the TSE are described in the following.

3.2.1 Encoder

The left part of Figure 1 is a hierarchical encoder which encodes the source content into context vectors by Gate Recurrent Unit (GRU) (Cho et al., 2014) or Transformer (Vaswani et al., 2017) (we take GRU as example in the following) and Graph Convolutional Network (GCN) (Kipf and Welling, 2017; Fu et al., 2019). More precisely, for the word encoder, we use GRU to capture the sequence dependency among words in the same sentence. As words with various POS play different roles in the sentence, we add the POS embedding to the corresponding word's embedding as Equation 1.

Moreover, with the aim of obtaining the inner relationships between different words in a sentence, we apply GCN to encode each word in every sentence separately with the dependency parsing results (Adj_{dep}), if a relation exists, the edge value between the two words is set to 1 in the adjacent matrix, 0 vice versa, the calculation process is shown in Equations 2-4.

$$I_{i,t} = Embed(w_{i,t}) + Embed(POS_{i,t}) \tag{1}$$

$$o_{i_{GRU}}, h_{i_{GRU}} = GRU(I_i) \tag{2}$$

$$o_{i_{GCN}} = GCN(o_{i_{GRU}}, Adj_{dep}). \tag{3}$$

$$o_i, h_i = GRU(o_{i_{GCN}}) \tag{4}$$

We use the similar modules for the sentence encoder compared to the word encoder while the weighted sum h_t^* of the re-located useful words' embedding vectors are used as the first GRU layer's inputs (described in 3.2.2). And we form the adjacent matrix for sentence GCN layer by the cosine similarity between any two sentences' TF-IDF vectors (Adj_{tf-idf}), the values in the matrix are converted to 1/0 by a certain threshold 0.3.

3.2.2 Relocating Critical Words

Different from previous work which used the encoder outputs of the last word as the corresponding sentence's embedding, we apply TSE to locate the corresponding critical words and then update the embedding of each sentence. Concretely, a attention layer is utilized to obtain the encoding input of sentence S_i at decoding step t. First, we use the hidden states h_i of the word-level encoder within sentence S_i and decoder hidden state s_t to calculate the similarity score e_w^t. Secondly, the $softmax$ function is used to calculate the attention weight a_w^t. The calculation process is shown in Equations 5-7,

$$e_i^t = v_w^t \tanh(W_{h,w}h_i + W_{s,w}s_t + b_{a,w}) \tag{5}$$

$$a^t = softmax(e^t) \tag{6}$$

$$h_t^* = \sum_i a_i^t o_i \tag{7}$$

where $v_w^t, W_{h,w}, W_{s,w}, b_{attn,w}$ are learnable parameters.

The weighted sum h_{t,S_i}^* of all sentences is transformed by GRU and GCN layers via Equations 2-4, the output vectors o_{S_i} are used as the representation of sentence S_i at decoding timestep t.

Similar to Equations 5-7, we adopt another attention layer to calculate the context vector $h_{t,s}^*$ based on dynamically updated sentence embeddings O_S. The context vector is applied to calculate the generation probability p_g and vocabulary distribution P_v by Equation 8 and Equation 9.

$$p_g^t = \sigma(W_{h^*}^{'} h_{t,s}^* + W_s^{'} s_t + W_y^{'} y_{t-1} + b^{'}) \tag{8}$$

$$p_v^t = softmax(W_{out}[s_t, h_{t,s}^*] + b) \tag{9}$$

Where $W_{h^*}^{'}, W_s^{'}, W_y^{'}, b^{'}, W_{out}, b$ are learnable parameters, y_{t-1} is the generated word at step $t-1$.

3.2.3 Decoder

For each decoding step, the key information from different parts of the source content which were picked by the TSE are utilized to generate the target word. In addition, the copy mechanism (Vinyals et al., 2015; Gu et al., 2016) is also adopted since some keywords in the title appeared in the source content. Different from the traditional copy mechanism with a pure sequence encoder, a hierarchical copy approach is proposed based on the hierarchical encoder framework. The attention weights over words within the same sentence and the sentence level attention scores are multiplied to calculate the copy probabilities for all words throughout the article content as shown in Equation 10. The final generation probability for word w at decoding step t over the extended vocabulary is calculated by Equation 11.

$$p_c^t = a_w^t * a_s^t \tag{10}$$

$$p^t(w) = p_g^t p_v^t(w) + (1 - p_g^t) \sum_{i:w_i = w} p_{c,i}^t \tag{11}$$

3.3 Acme Coverage

As mentioned above, seq2seq models with copy mechanisms usually suffer from the repetition problem. Especially in the title generation tasks, the repeated words are conspicuous in a concise title and resulting in poor readability. The commonly used Vanilla Coverage (VC) mechanism maintains a coverage vector cov^t to calculate the sum of attention distribution over all previous decoder steps. But for the title generation scenario, the copied word may appear in different positions with different attention weights, the vanilla coverage mechanism only penalizes the model to avoid repeatedly attending to the same locations but not the same word. Besides, since the attention weights in our model are in a hierarchical mode, the weights' values of closely located words may be similar sometimes. The vanilla coverage mechanism may wrongly penalize words that have not been generated actually but are essential in the following steps.

With the aim of making up for the two shortcomings of the vanilla coverage mechanism, we proposed a novel mechanism named Acme Coverage, which only sums over all the attention weights for the words generated at each decoding step if they appeared in the content actually. The final generation probability p^t over the extended vocabulary is used to select the truly generated word w^t, and all the attention weights corresponding to w^t in a_w^t will be added together for loss penalization. The calculation is shown in Equations 12 and 13, where $I_{a_w^t}$ is an indicator function.

$$cov^t = \sum_{t=0}^{t-1} I_{a_w^t} * a_w^t \tag{12}$$

$$I_{a_w^t} = \begin{cases} 1 & i == argmax(p^t) \\ 0 & others \end{cases} \tag{13}$$

Considering the AC mechanism, the calculation of Equation 5 of word-level attention weight will be changed to Equation 14.

$$e_i^t = v^t \tanh(W_h h_i + W_s s_t + W_c cov_i^t + b_a) \tag{14}$$

Similar to the work in (See et al., 2017), we also add the coverage loss in the final loss function to enhance the model's ability to accurately copy while avoiding duplication. The composite loss function is shown in Equation 15

$$L_t = -\log(p^t(w)) + \lambda \sum_i \min(a_{w,i}^t, cov_i^t) \tag{15}$$

4 Experiments

4.1 Datasets

We conduct sufficient experiments over two publicly available title generation datasets with brief but informative titles to verify the effectiveness of our method, **AESLC** (Zhang and Tetreault, 2019), and **Chinese Gigaword** (Parker et al., 2011).

AESLC The AESLC is an annotated email subject line corpus which is a collection of email messages of employees in Enron Corporation. The average title length in AESLC is 4 words, much shorter than other summary generation tasks. There are 18k samples in AESLC, with train/val/test: 14,436/1,960/1,906. Notably, AESLC has two different targets in validation and test set, one is the original subject of the email, another is annotated subjects from three annotators according to the content.

Chinese Gigaword This dataset is a collection of Chinese news articles, it contains paragraphs from the Chinese Gigaword Fifth Edition release. The average lengths of the source articles and target titles are 665.5 and 16.3. In total there are more than 5M articles, and we sample 74689 for training and 9400/9345 for validation and testing.

4.2 Evaluation

To evaluate the performance of our proposed model and compare it with other baselines, we use the automatic metrics from text summarization and machine translation: ROUGE 1/2/L (Lin and Och, 2004) and METEOR (Denkowski and Lavie, 2014) to measure the quality of the generated titles.

Besides, human evaluation is also adopted to evaluate the quality of the generated news headlines in three dimensions. The first one is the fluency which indicates whether the title is grammatical correct and in high readability. The second one is to measure the relevance between the generated title and the input article. The last but most important one is the usability of the generated title which means whether the brief title can be used as a formal one in practical scenarios. The scores for fluency and relevance are between 1-5. A higher score means the quality is better. The usability is a simple 0/1 judgment. The average score of fluency, relevance and the available ratio are used to compare the quality of the generated titles by different settings of the coverage mechanism.

4.3 Implementation Details

Our model is with 2 layers of bidirectional GRU (or Transformer) and GCN for word and sentence encoders respectively. The POS of each word and dependency parsing information are obtained by the

methods	DEV				TEST			
	R-1	R-2	R-L	METEOR	R-1	R-2	R-L	METEOR
PG-net (See et al., 2017)	18.02	5.73	16.63	10.83	17.02	5.45	15.78	10.31
Zhang and Tetreault (2019)	25.41	11.34	25.07	9.83	23.67	10.29	23.44	9.37
T5 (Raffel et al., 2020)	23.74	11.73	23.43	8.67	23.68	11.97	23.27	8.92
SimCLS (Liu and Liu, 2021)	25.67	**12.36**	25.42	9.72	24.52	12.35	24.03	9.63
PEGASUS-base (Zhang et al., 2020)	-	-	-	-	34.85	18.94	34.10	-
PEGASUS-large (Zhang et al., 2020)	-	-	-	-	**37.69**	**21.85**	**36.84**	-
Human Annotation	23.43	9.71	22.17	10.87	23.90	10.09	22.75	11.04
TSE-AC	26.28	11.53	25.73	**11.17**	24.91	10.91	24.27	11.09
TSE-AC-Trans	**26.35**	12.07	**25.99**	11.08	**25.18**	11.86	**24.59**	**11.25**

Table 2: The performance against the original subject of AESLC. The top two lines are results referred from (Zhang and Tetreault, 2019). The T5 and SimCLS are results conducted by ourselves with models proposed recently. Human Annotation means using annotated subjects from the third annotator as predict results. TSE-AC is our model with GRU layers while TSE-AC-Trans use Transformer layers instead.

methods	DEV				TEST			
	R-1	R-2	R-L	METEOR	R-1	R-2	R-L	METEOR
PG-net	23.37	7.36	20.99	16.27	23.31	7.28	20.83	15.68
Zhang and Tetreault (2019)	25.39	10.94	24.72	13.04	26.11	11.43	25.64	13.52
Original Subject	24.38	10.15	23.00	16.49	25.47	10.40	23.15	14.08
T5 (Raffel et al., 2020)	21.31	10.26	20.88	11.84	21.76	10.80	21.33	11.92
SimCLS (Liu and Liu, 2021)	22.14	10.03	21.25	12.33	22.08	10.82	21.57	12.40
TSE-AC	29.55	12.52	28.09	15.87	29.44	12.41	28.20	15.53
TSE-AC-Trans	**30.01**	**13.29**	**29.13**	**16.55**	**30.20**	**13.32**	**29.07**	**16.17**
Human Annotation	35.93	17.76	33.55	21.74	36.19	17.75	33.50	21.42

Table 3: The performance against two human annotations as references of AESLC. Original Subject means using the original subjects as predict results.

Spacy toolkit (Honnibal et al., 2020). The threshold of TF-IDF cosine similarity used in sentence-level adjacent matrix is 0.3. The decoder includes two layers of GRU (or Transformer) and one fully connected layer with vocabulary size 35000. The embedding size of words and POS are 256 as same as the hidden size. The Adam (Kingma and Ba, 2014) optimizer is used to train the model with learning rate and dropout ratio set to 0.0005 and 0.5. The coverage loss is added after 2 epochs of training steps with weight set to 1.0. The training is converged in 6 epochs for AESLC dataset and 20 epochs for Gigaword with one NVIDIA P100 GPU and batch size set to 8, the maximum value of gradient clip is 5.0. In the model prediction stage, we use beam search with size 4 to generate the titles for samples in the test sets. The code is available at https://github.com/alipay/Timestep-aware-SentenceEmbedding-and-AcmeCoverage.

4.4 Results

Automatic Evaluation The results of experiments over AESLC are given in Table 2 and Table 3. In Table 2, where the original subjects are ground-truth, our model achieves the best results on all automatic evaluation metrics for both validation and test sets except Rouge-2 and the PEGASUS (Zhang et al., 2020) models. It needs to be acknowledged that PEGASUS performances are much better since it is pre-trained specifically for summarization tasks. We think it is not fair to compare our model with it. Different from the unilateral improvement in (Zhang and Tetreault, 2019), our method not only obtains about **1%** (abs.) improvements on ROUGE-1/L scores but also significantly surpasses the performance on the METEOR score with around **2%** (abs.). Previous works generate titles that only focus on limited sentences or static sentences' embeddings more often produce trivial words, instead, our method generates each target word of the title meticulously via TSE to re-locate the related words and update the sentences' embeddings. Thus, with the increase of ROUGE score, our method also obtains improvement of METEOR score. Identically, when the ground-truth is the human-annotated subjects, our method outperforms competitive baselines remarkably, the results are summarized in Table 3. The improvement on ROUGE-1/L is over **3%** (abs.) and about **2%** (abs.) on METEOR score. We also achieved the best Rouge-2 score in this setting with about **2%** improvements. Besides, we also replace the GRU layers with Transformer layers to test the usefulness of TSE with different backbone and the performance is a little better. It demonstrates that the two novel mechanisms, TSE and AC are effective for different kinds of neural structures.

methods	R-1	R-2	R-L	METEOR
PG-net	50.25	43.36	49.84	27.46
PALM	58.30	49.49	57.69	32.45
SimCLS	59.29	50.12	58.01	32.86
TSE-VC	58.02	49.30	57.15	31.51
TSE-AC	62.17	53.09	61.21	**36.24**
TSE-VC-Trans	59.11	50.06	58.21	32.44
TSE-AC-Trans	**63.09**	**54.13**	**61.97**	36.17

Table 4: The performance comparison on Gigaword. TSE-VC means using vanilla coverage with TSE.

It is worthy to indicate that our small model achieves much better results compared with the classical pre-trained model T5 (Raffel et al., 2020) and the SimCLS which use the results generated by T5 as candidates to further training the model with contrastive learning (Liu and Liu, 2021). We do not compared to the customized pre-trained model for summarization in (Zhang et al., 2020) since it is unfair. The improvements are higher when using human annotation as references. Through in-depth analysis of the generated results, we found that the pre-trained model suffers from aligning with longer inputs and the short titles with words scattered among the whole content. This information aggregation process has a large gap compared to the pre-training tasks. And the pre-trained T5 without coverage mechanism is poor at tackling the repetition problem. In addition, the pre-trained model with much more parameters and general knowledge learns more easily what it observed, so the performance in Table 3 degrades heavily when the ground-truth is different from the learning target.

Moreover, we also implement experiments over Gigaword in Chinese with longer source articles and titles, the results demonstrate similar conclusions. Specifically, we adopt PALM (Bi et al., 2020b) as a baseline on this dataset. PALM is a pre-trained generation model and we obtained the pre-trained model with a large Chinese corpus from the author, it performs better on many natural language processing tasks in Chinese than T5. Our model achieves over **10%** (abs.) higher score on Gigaword with PG-net as baseline and about **5%** (abs.) improvements over PALM. It indicates the versatility of our approach that is capable for different scenarios of title generation.

Human Evaluation We only conduct human evaluation on Chinese Gigaword dataset since the three human annotators are with Chinese as native language. 200 news texts are sampled from the corpus randomly. The comparison among SimCLS and different settings of the coverage mechanism with our proposed approach are shown in Table 5, from which we can conclude that the general performance of our method with TSE and coverage mechanism can provide titles with high quality. Though our model without coverage mechanism performs worse than SimCLS, two kinds of coverage mechanisms can improve the performance effectively. Especially the AC can raise the titles' quality to higher level and is acceptable to be used in practical scenarios with over **80%** of the generated titles being usable for real news.

method	Fluency	Relevance	Usability
SimCLS	3.07	3.38	65.32%
w/o Cov	2.78	3.16	50.12%
D-VC	3.15	3.56	68.31%
D-AC	**3.54**	**3.72**	**80.27%**

Table 5: Human evaluation on fluency, relevance and usability with different coverage settings.

Analysis of Computation Complexity We further analyze the computation complexity of our model and compared it to traditional seq2seq models. Assuming that the input content has M sentences, and each sentence has N words averagely, the target title contains T tokens. The time complexity of vanilla seq2seq models like PG-net is $O(M * N) + O(T)$ since they have to encode the input content sequentially and generate the title word by word. For our TSE-AC model, the encoder with hierarchical architecture can parallel encode words in each sentence, and we need to update the embedding of each sentence for every target word, so the time complexity is $O(N) + O(T * M)$. As we mentioned before, we focus on short title generation tasks in this paper which means $T < N$, and it can be derived that $O(M*N)+O(T) >= O(M*(T+1))+O(T) = O(T*M)+O(M)+O(T) > O(M)+O(T*M)$. As a result, if we split the input content into M sentences with N words in each sentence and make sure $M < N$, our model's computation complexity is less than traditional seq2seq models. When change the backbone model from GRU to Transfomers or pure hierarchical encoder, our model is a little complex since transformer layers based encoder is parallel naturally, but the cost of re-computation of sentences' embeddings are accept-

Ref.	Condition	DEV				TEST			
		R-1	R-2	R-L	METEOR	R-1	R-2	R-L	METEOR
Ori	Our Best	**26.35**	**12.07**	**25.99**	**11.08**	**25.18**	**11.86**	**24.59**	11.25
	w/o TSE	24.25	9.47	23.64	10.03	22.76	8.68	22.30	10.05
	w/o WG	25.34	10.26	24.92	10.95	24.87	8.82	23.99	11.21
	w/o SG	24.62	8.79	24.18	9.42	23.17	7.93	22.80	**13.14**
	w/o POS	23.30	8.54	22.82	9.48	23.02	8.76	22.69	9.72
	w/o Cov	25.02	9.64	24.39	10.35	23.38	8.68	22.83	9.77
	w/ VC	25.62	10.36	25.18	11.03	24.11	10.17	23.55	10.45
HA	Our Best	**30.01**	**13.29**	**29.13**	**16.55**	**30.20**	**13.32**	**29.07**	**16.17**
	w/o TSE	23.63	9.05	22.82	12.78	24.69	9.29	23.90	13.24
	w/o WG	25.20	10.31	24.60	13.42	26.42	9.76	25.41	14.21
	w/o SG	25.53	9.73	24.75	9.09	26.62	10.05	25.96	13.49
	w/o POS	24.74	9.33	23.99	13.13	25.55	9.70	24.90	13.34
	w/o Cov	26.71	9.91	25.70	13.24	26.90	9.82	25.98	13.93
	w/ VC	27.40	11.02	26.51	14.23	27.28	11.12	26.50	14.25

Table 6: The ablation study of our method on AESLC. Ori and HA indicate original references and human annotations. w/o TSE represents the model without TSE which means the sentences' embeddings are frozen for all decoder steps, w/o WG and w/o SG mean that the inner relations among words or sentences are not used. POS means part of speech information and Cov indicates coverage mechanism. w/ VC means using vanilla coverage.

Coverage	PG-net				TSE			
	R-1	R-2	R-L	METEOR	R-1	R-2	R-L	METEOR
w/o Cov	21.65	6.71	19.47	10.35	26.90	9.82	25.98	13.93
w/ VC	24.09	7.11	22.52	13.96	27.28	11.12	26.50	14.25
w/ AC	**26.50**	**10.08**	**25.08**	**15.46**	**29.44**	**12.41**	**28.20**	**15.53**

Table 7: The performance comparison among different settings of the coverage mechanism for title generation by PG-net and our TSE-AC on AESLC test set with annotated subjects as reference. w/o Cov means coverage mechanism not used. The results of PG-net are reproduced by ourselves using publicly available code.

able and meaningful since the titles are always brief.

4.5 Ablation Study

Since there are several components in our proposed model, including TSE mechanism, GCN layer for inner relations encoding, POS information, and coverage mechanism, we conduct comprehensive experiments to discover whether the specific part gives positive influence to the model's performance or negative. The results are shown in Table 6.

With the original email subjects as references, the performances declined heavily on ROUGE score when the TSE is discarded and the value of METEOR reduced most without the POS information as model's inputs. The performances on human annotations have the same phenomenon. Overall, the absence of these two components has the greatest negative impact on the performance of our model. The results demonstrate that the TSE mechanism is essential for title generation since it can re-locate the key information and update the sentences' embeddings at each decoding timestep. The POS information is also useful because words with different POS play a different role in the whole texts, which facilitates the TSE to focus on the corresponding critical words. For the effectiveness of GCN for inner relations encoding, the results indicate that GCN encoder layers have positive influence on ROUGE obviously but inconsistent impact on METEOR metric while the value is much higher when sentence level GCN is abandoned with original subject as a reference. This may result from the inconsistency of the original subject and the annotations. The coverage mechanism is also useful especially our novel AC. The ablation study demonstrates that the TSE and AC is effective for brief title generation tasks.

To compare the effectiveness of our AC mechanism with the vanilla one directly, we conduct experiments with different coverage settings on our proposed model and PG-net. The results shown in Table 4,6 and 7 indicate that our novel AC can achieve better performances not only with TSE but also other title generation methods or backbones. And it means that AC is more general and can make up for the shortcoming of VC. From the second term of Equation 15, the coverage loss will make the model ignore the words which have been pay more attention before as much as possible. If the word with high attention in previous timestep and has not been generated actually, the models with

VC mechanism are easier to ignore it. And our AC mechanism can well avoid this wrong punishment by only considering the actually generated words. The experiments also indicate that the combination of TSE and AC results in higher improvements compared to changing the backbone from GRU to Transformer.

4.6 Case Study

In Table 8 (see Appendix A), some examples with subjects and corresponding email contents are given to demonstrate the ability of our novel model. From the first three samples, we find that our model can generate more accurate subjects compared to the baseline method proposed in (See et al., 2017), of which the subjects need to be concluded based on the entire source content, not just the first or last few sentences. And the related words are different within the same sentences for each target word according to their location and inner relations. The traditional seq2seq model ignores the inner relationship among words and sentences but only encodes their sequential information. The work in (Zhang and Tetreault, 2019) first extracts some key sentences and generated subjects merely base on those sentences will lose many useful clues to generate more valuable keywords. Moreover, although our model is trained with original subjects, it can generate more reasonable targets compared to the original ones in most cases.

The last two examples indicate that our model can obtain more suitable subjects even if the key information is contained in the first few sentences. As the title is a highly condensed summary of the article, each word in the title needs to be confirmed based on different parts of the original content. Our model with TSE and AC mechanisms can finely relocate the valuable words scattered in the original text, and the selected words are used to update the sentences' embeddings for each decoding timestep. The context vector is further modified so as to better summarize the input article.

All the examples show that our novel AC mechanism can obtain subjects with more information while the VC lost some key information more or less. Our model TSE-AC can generate titles with more complete and accurate information compared to PG-net, and the repetition problem is mitigated while preserving more critical words compared with T5 and vanilla coverage.

5 Conclusion

In this paper, we proposed a novel title generation framework with Timestep aware Sentence Embedding, which re-locates critical words dynamically for each target word from the source content based on the decoding states to update the corresponding sentence's embedding. Moreover, we present Acme Coverage that can accurately penalize the probability of the word which has been generated actually. The experiments demonstrate that our approach achieves the state of the art performance on different kinds of title generation scenarios.

6 Acknowledgement

This work is supported by the Postdoctoral Program of Zhejiang Province. Thanks for the guidance and help from my advisors Prof. Wei Chu (`weichu.cw@antgroup.com`) and Prof. Huajun Chen (`huajunsir@zju.edu.cn`). Thanks to my supervisor Mr. Taifen Wang (`taifeng.wang@antgroup.com`) of the invaluable suggestions for this work.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bin Bi, Chenliang Li, Chen Wu, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2020a. PALM: pre-training an autoencoding&autoregressive language model for context-conditioned generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8681–8691. Association for Computational Linguistics.

Bin Bi, Chenliang Li, Chen Wu, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2020b. PALM: pre-training an autoencoding&autoregressive language model for context-conditioned generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8681–8691. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. Graphrel: Modeling text as relational graphs for joint entity and relation extraction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1409–1418. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. *arXiv preprint arXiv:1808.10792*.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020. Hooks in the headline: Learning to generate headlines with controlled styles. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5082–5093. Association for Computational Linguistics.

Rong Jin and Alexander G. Hauptmann. 2001. Automatic title generation for spoken broadcast news. In *Proceedings of the First International Conference on Human Language Technology Research, HLT 2001, San Diego, California, USA, March 18-21, 2001*. Morgan Kaufmann.

Rong Jin and Alexander G. Hauptmann. 2002. A new probabilistic model for title generation. In *19th International Conference on Computational Linguistics, COLING 2002, Howard International House and Academia Sinica, Taipei, Taiwan, August 24 - September 1, 2002*.

Paul E. Kennedy and Alexander G. Hauptmann. 2000. Automatic title generation for EM. In *Proceedings of the Fifth ACM Conference on Digital Libraries, June 2-7, 2000, San Antonio, TX, USA*, pages 230–231. ACM.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Chenliang Li, Weiran Xu, Si Li, and Sheng Gao. 2018. Guiding generation for abstractive text summarization based on key information guide network. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 55–60. Association for Computational Linguistics.

Haoran Li, Junnan Zhu, Jiajun Zhang, Chengqing Zong, and Xiaodong He. 2020a. Keywords-guided abstractive sentence summarization. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8196–8203. AAAI Press.

Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du. 2020b. Leveraging graph to improve abstractive multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6232–6243. Association for Computational Linguistics.

Chin-Yew Lin and FJ Och. 2004. Looking for a few good metrics: Rouge and its evaluation. In *Ntcir Workshop*.

Wenfeng Liu, Peiyu Liu, Yuzhen Yang, Jing Yi, and Zhenfang Zhu. 2019. A< word, part of speech> embedding model for text classification. *Expert Systems*, 36(6):e12460.

Yixin Liu and Pengfei Liu. 2021. Simcls: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021,* *(Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 1065–1072. Association for Computational Linguistics.

Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage embedding models for neural machine translation. *arXiv preprint arXiv:1605.03148*.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Robert Parker, D Graff, K Chen, J Kong, and K Maeda. 2011. Chinese gigaword fifth edition (ldc2011t13). *Linguistic Data Consortium*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018a. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018b. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Baskaran Sankaran, Haitao Mi, Yaser Al-Onaizan, and Abe Ittycheriah. 2016. Temporal attention model for neural machine translation. *arXiv preprint arXiv:1608.02927*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. From neural sentence summarization to headline generation: A coarse-to-fine approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4109–4115. ijcai.org.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. *arXiv preprint arXiv:1601.04811*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2692–2700.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1480–1489. The Association for Computational Linguistics.

Bowen Yu, Xue Mengge, Zhenyu Zhang, Tingwen Liu, Wang Yubin, and Bin Wang. 2020. Learning to prune dependency trees with rethinking for neural relation extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3842–3852.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Rui Zhang and Joel R. Tetreault. 2019. This email could save your life: Introducing the task of email subject line generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 446–456. Association for Computational Linguistics.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2205–2215. Association for Computational Linguistics.

A Generated Examples Appendix

email content: We have received the confidentiality agreement with Novo media group inc dated May 15th. Copies will be distributed to Dave Samuels and Bob Shults. I am also attaching an updated list of the enrononline confidentiality agreements.
original subject: Novo media group list of confidentiality agreements
annotated subject: current and updated confidentiality agreements
generated subject by PG-net: confidentiality agreement
generated subject by T5: confidentiality agreement
generated subject by TSE-VC: updated list confidentiality
generated subject by TSE-AC: updated list of confidentiality agreements

email content: We are planning a party it has been a while since the group has had a party and my daughter is going to be a year old. So we are planning a party for November 14th and Vandhana and I would like to invite everyone in research and their family. ...
original subject: birthday party
annotated subject: birthday party invitation
generated subject by PG-net: ces year
generated subject by T5: party
generated subject by TSE-VC:party planning
generated subject by TSE-AC:party invite

email content: eSource Presents Lexis-Nexis Training Basic Lexis-Nexis Basic is geared to the novice or prospective user. You will learn the basics of getting around Nexis.com... Attend our Lexis-Nexis Basics Clinic: November 6 1:00-2:00 PM EB572 Due Diligence... Attend our Lexis-Nexis Due Diligence Clinic: November 6 2:30 - 4:00 PM EB572. Seats fill up fast! To reserve a seat, please call Stephanie E. Taylor at 5-7928...Source presents free Lexis-Nexis Online Training. ...
original subject: Lexis-Nexis Training: Houston & Worldwide / Dow Jones Training
annotated subject: online training clinic for lexis-nexis
generated subject by PG-net: training for lexis
generated subject by T5: lexis-nexis-neixs-neixs
generated subject by TSE-VC: lexis online training
generated subject by TSE-AC: lexis nexis online training

email content: I always compile a contact list for energy operations during the holidays thanksgiving Christmas and new years. Just let me know who appropriate contacts will be especially for the DPR and MPR during the dates that you are out. ...
original subject: vacation plans
annotated subject: energy operations contact list
generated subject by PG-net: contact list
generated subject by T5: holiday contact list
generated subject by TSE-VC:contact info energy
generated subject by TSE-AC:contact info for energy

email content: Please find attached the latest and what should be the final for the immediate period of time copy of the marketing list. Please filter the pa column by your name to double check against the list you are currently working off of there are some smaller subsids of larger companies previously assigned now listed. ...
original subject: latest marketing list
annotated subject: marketing list
generated subject by PG-net: marketing list
generated subject by T5: enron wholesale markets list
generated subject by TSE-VC: check the marketing
generated subject by TSE-AC: check the marketing list

Table 8: Examples of email subject generated by our model and other baselines