Multitasking Framework for Unsupervised Simple Definition Generation

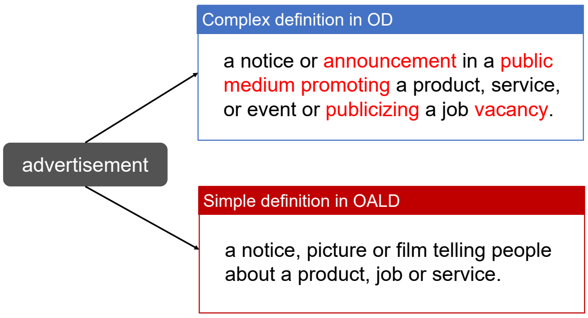
Cunliang Kong[134], **Yun Chen**[2], **Hengyuan Zhang**[1], **Liner Yang**[134][*], **Erhong Yang**[134]
[1]School of Information Science, Beijing Language and Culture University
[2]School of Information Management & Engineering,
Shanghai University of Finance and Economics
[3]National Language Resources Monitoring and Research Center Print Media Branch,
Beijing Language and Culture University
[4]Beijing Advanced Innovation Center for Language Resources,
Beijing Language and Culture University

Abstract

The definition generation task can help language learners by providing explanations for unfamiliar words. This task has attracted much attention in recent years. We propose a novel task of Simple Definition Generation (SDG) to help language learners and low literacy readers. A significant challenge of this task is the lack of learner's dictionaries in many languages, and therefore the lack of data for supervised training. We explore this task and propose a multitasking framework **SimpDefiner** that only requires a standard dictionary with complex definitions and a corpus containing arbitrary simple texts. We disentangle the complexity factors from the text by carefully designing a parameter sharing scheme between two decoders. By jointly training these components, the framework can generate both complex and simple definitions simultaneously. We demonstrate that the framework can generate relevant, simple definitions for the target words through automatic and manual evaluations on English and Chinese datasets. Our method outperforms the baseline model by a 1.77 SARI score on the English dataset, and raises the proportion of the low level (HSK level 1-3) words in Chinese definitions by 3.87% [1].

1 Introduction

Helping language learners understand words in doubt is an important topic in the field of Intelligent Computer-Assisted Language Learning (ICALL) (Segler et al., 2002; Enayati and Gilakjani, 2020; Lolita et al., 2020). In recent years, researchers attempted to automatically generate definitions for words rather than formulating predefined word-definition inventories (Ishiwatari et al., 2019; Yang et al., 2020; Huang et al., 2021). There are two reasons for this. Firstly, it can be difficult for users to distinguish which sense is appropriate in the current context because of the cognitively inaccurate nature of discrete sense boundaries (Rosch and Mervis, 1975; Kilgarriff, 1997; Tyler and Evans, 2001). Secondly, the predefined inventories need to be updated manually by lexicographers, which is time-consuming and causes dictionaries to lag behind the ever-changing language usage.



a notice or announcement in a public medium promoting a product, service, or event or publicizing a job vacancy.

a notice, picture or film telling people about a product, job or service.

Figure 1: Different definitions for *advertisement* in the Oxford Dictionary (OD) and Oxford Advanced Learner's Dictionary (OALD).

Different from previous work (Noraset et al., 2017; Gadetsky et al., 2018; Mickus et al., 2019; Kong et al., 2020) that focused only on how to generate definitions, we further propose a novel task of **S**imple **D**efinition **G**eneration (SDG). Making the definitions easier to read and understand could benefit the language learners, low literacy readers, as well as helping people with aphasia or dyslexia. For example, compared with the Oxford Dictionary (OD), the Oxford Advanced Learner's Dictionary (OALD) has simpler definitions, which are specifically designed for language learners. As shown in Figure 1, the definition of the word *advertisement* in OALD does not contain difficult words or phrases such as *announcement* and *public medium*.

The goal of SDG task is to generate simple definitions for languages that lack learner's dictionary. For example, Chinese as Second Language (CSL) learners do not have suitable dictionaries. As Zhang (2011) pointed out, since the difficulty of

[*] Corresponding author

[1] Code can be found at https://github.com/blcuicall/SimpDefiner.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5934 - 5943
May 22-27, 2022 ©2022 Association for Computational Linguistics

definitions is not considered, the existing dictionary cannot meet CSL learner's needs.

The SDG task is challenging because it requires a model to learn from a standard dictionary containing complex definitions and then generate simple ones, and hence fully unsupervised. A seemingly feasible solution is to generate definitions first and then simplify them, i.e., the generation-simplification pipeline. However, the simplification task requires dataset with complex-simple sentence pairs, and such data is also difficult to find in languages other than English (Martin et al., 2020). Besides, the pipeline methods do not perform well due to accumulated errors (Section 6.1).

To solve this dilemma and bridge the gap between practical needs for simple definitions and current trivial definition generation systems, we present a novel method for the SDG task. As illustrated in Figure 2, our method leverages a multitasking framework **SimpDefiner** to generate simple definitions by performing three sub-tasks at the same time, which are definition generation, text reconstruction, and language modeling tasks. The framework consists of a fully shared encoder and two partially shared decoders. We disentangle the complexity factors from the text by designing a parameter sharing scheme. Particularly, we share parameters in Complexity-Dependent Layer Normalization and Complexity-Dependent Query Projection of the transformer architecture (Vaswani et al., 2017) to control the complexity (Section 3.3). Through joint learning and sharing parameters between the decoders, the SimpDefiner is able to generate complex and simple definitions simultaneously.

Main contributions of our paper are listed below:

- For the first time, we propose the task of SDG to generate simple definitions without supervised training data.

- We propose a multitasking framework SimpDefiner to tackle this task. Through joint training three sub-tasks, the framework can generate complex and simple definitions simultaneously.

- Both automatic and manual evaluations demonstrate the effectiveness of SimpDefiner. The framework outperforms the baseline model by 1.77 SARI score on the English test set. And the proportion of low level words (HSK level 1-3) in generated definitions raised by 3.87% on the Chinese test set.

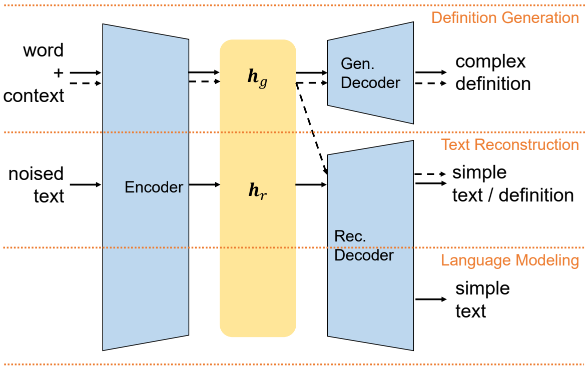


Figure 2: The SimpDefiner consists of three sub-tasks. Gen. means generation and Rec. means reconstruction. The solid black lines indicate the data-flow during training, and the dashed black lines indicate the data-flow during inference.

2 Related Work

2.1 Definition Generation

The definition generation task is first introduced by Noraset et al. (2017). Although this task is proposed as a potentially useful tool for explainable AI, many subsequent works believe that it can assist language learning by giving definitions for words in the text (Ishiwatari et al., 2019; Mickus et al., 2019; Yang et al., 2020).

Various studies attempted to generate multiple different definitions for polysemous words. Gadetsky et al. (2018) tackled this problem by computing the AdaGram vectors (Bartunov et al., 2016) of input words, which are capable of learning different representations at desired semantic resolutions. However, generating different definitions based on contexts, i.e., example sentences, became the mainstream method (Chang et al., 2018; Reid et al., 2020; Li et al., 2020; Bevilacqua et al., 2020). Among them, some studies used pre-trained language models to obtain contextualized embeddings. Reid et al. (2020) initialized encoders with BERT (Devlin et al., 2019) and employed variational inference for estimation and leveraged contextualized word embeddings for improved performance. Bevilacqua et al. (2020) employed a novel span-based encoding scheme to fine-tune a pre-trained English encoder-decoder system to generate definitions. Huang et al. (2021) leveraged the T5 (Raffel et al., 2019) model for this task and introduced a re-ranking mechanism to model specificity in definitions.

Our proposed SimpDefiner also takes the given word and context as input. Differently, our main focus is to generate definitions with appropriate complexity to better help language learners. Besides, our model is based on MASS (Song et al., 2019), which is a pre-trained encoder-decoder model and is suitable for generation tasks.

2.2 Sentence Simplification

Researchers usually regard the sentence simplification task as a monolingual variant of machine translation (MT) (Wubben et al., 2012). Benefiting from the advancement of neural machine translation, this task has also made great progress in recent years.

Lately, many works built upon the Seq2Seq MT model (Sutskever et al., 2014) performed well. First attempted by Nisioi et al. (2017), the Seq2Seq models for this task are able to perform lexical simplification and content reduction simultaneously by training on complex-simple sentence pairs. This method was inherited and improved by many subsequent works, such as combining with the reinforcement learning method by setting a simplification reward (Zhang and Lapata, 2017), augmenting memory capacities (Vu et al., 2018) or training with multitasking on entailment and paraphrase generation (Guo et al., 2018). Martin et al. (2019) proposed to prepend additional prompt tokens to source sentences at train time, which enables the end-users to condition the simplifications returned by the model on attributes like length, lexical complexity, and syntactic complexity. This controllable simplification system (called ACCESS) and its improved version MUSS (Martin et al., 2020) achieved SOTA results on the Turk corpus in terms of the SARI metric (Xu et al., 2016).

The generation-simplification pipeline methods are used as baselines of the SDG task, and we use both ACCESS and MUSS models for the simplification. Unlike the baseline, the SimpDefiner can generate simple definitions directly, alleviating the accumulated errors.

2.3 Unsupervised Style Transfer

Style transfer aims to change the style attributes while preserving the content. Our work is related to unsupervised style transfer by regarding the text complexity as one of the style attributes (Kawashima and Takagi, 2019). Dumoulin et al. (2017) demonstrated that the neural networks can capture the artistic style of a diversity of paintings. The authors discovered that adjusting parameters in the layer normalization mechanism leads to different artistic styles. This method permits users to transform images to arbitrary styles learned from individual paintings. Jin et al. (2020) successfully applied this method to the task of headline generation, allowing the model to generate headlines of a specific style, such as humorous, romantic or click-baity, in an unsupervised manner.

By treating the task of simplification as a variant of style transfer, we borrow the insight of learning complexity-dependent parameters in the Layer Normalization mechanism. Additionally, we introduce the language modeling task into SimpDefiner, which is to enhance the decoder and make it more sensitive to text complexity.

3 Method

We integrate three sub-tasks of definition generation, text reconstruction, and language modeling into the SimpDefiner. This section first gives a formal definition of the SDG task, then introduces each sub-task, and finally the parameter sharing scheme.

3.1 Task Formulation

The SDG task is to generate a simple definition $\boldsymbol{d}^{sim}$ for a given word and context $(w^*, \boldsymbol{c})$, where $\boldsymbol{c} = [w_1, \ldots, w^*, \ldots, w_n]$ is a sentence containing w^*. This task is challenging because there is no corpus like $\{(w_i^*, \boldsymbol{c}_i, \boldsymbol{d}_i^{sim})\}_{i=1}^N$ and hence it is fully unsupervised.

The only data available in this work include a standard dictionary dataset $G = \{(w_i^*, \boldsymbol{c}_i, \boldsymbol{d}_i^{com})\}_{i=1}^N$ and a simple text corpus $Y = \{\boldsymbol{y}_i\}_{i=1}^M$. Note that we use $\boldsymbol{d}^{com}$ for complex definitions and $\boldsymbol{d}^{sim}$ for simple ones.

3.2 Multitasking Framework

We design the three sub-tasks in the SimpDefiner to learn different abilities. Cooperating with each other, the entire framework obtains the ability to compute the conditional probability $P(\boldsymbol{d}^{sim}|w^*, \boldsymbol{c})$ of simple definitions in a zero-shot manner.

Specifically, the definition generation task aims to model the probability of a complex definition given the word and context $P(\boldsymbol{d}^{com}|w^*, \boldsymbol{c})$ (Section 3.2.1). And the text reconstruction task aims to model the probability of a simple sentence given the corrupted version $P(\boldsymbol{y}|\tilde{\boldsymbol{y}})$ (Section 3.2.2). As we can see, neither task can directly get the $P(\boldsymbol{d}^{sim}|w^*, \boldsymbol{c})$. To solve the problem, we assume that complexity and semantic information are controlled by different parameters in the decoders, and we attempt to disentangle the complexity factors from the text through a carefully designed parameter sharing scheme. In the inference stage, we obtain a simple definition by feeding the encoded hidden state into the reconstruction decoder as in Figure 2. The detailed parameter sharing scheme is in Section 3.3.

Nevertheless, the complexity information may still be kept in some shared parameters, resulting in the reconstruction decoder fail to generate simple definitions occasionally. Eliminating the complexity information in all shared parameters is obviously technically impossible. Instead, we introduce the language modeling task (Section 3.2.3) to enhance the reconstruction decoder and make it more *focused* on simple text generation. The experiment results in Section 6 confirm our assumption.

3.2.1 Definition Generation Task

We follow the mainstream method (Yang et al., 2020; Kong et al., 2020; Reid et al., 2020) to concatenate the word and context together with a special token [SEP] as $\mathrm{x} = (w^*; \text{[SEP]}; \boldsymbol{c})$. The entire sequence is then fed into SimpDefiner, and the definition is obtained by the following language model:

$$P(\boldsymbol{d}^{com}|\mathrm{x}; \boldsymbol{\theta}_g) = \prod_t P\left(\boldsymbol{d}_t^{com}|\boldsymbol{d}_{<t}^{com}, \mathrm{x}; \boldsymbol{\theta}_g\right), \quad (1)$$

where $\boldsymbol{d}_t^{com}$ is the t-th token of the definition, and $\boldsymbol{\theta}_g$ is the set of parameters. The model is optimized using the following loss function.

$$\mathcal{L}_{gen}(\boldsymbol{\theta}_g) = -\sum_{(\mathrm{x}, \boldsymbol{d}^{com}) \in G} \log P(\boldsymbol{d}^{com}|\mathrm{x}; \boldsymbol{\theta}_g) \quad (2)$$

3.2.2 Text Reconstruction Task

We corrupt each sentence in the corpus Y by randomly deleting or blanking some words and shuffling the word orders. And then we obtain a new corpus $\tilde{Y} = \{(\tilde{\boldsymbol{y}}_i, \boldsymbol{y}_i)\}_{i=1}^M$, and $\tilde{\boldsymbol{y}}$ is a corrupted version of $\boldsymbol{y}$. We input $\tilde{\boldsymbol{y}}$ into SimpDefiner and obtain $\boldsymbol{y}$ by solving a self-supervised task of

$$P(\boldsymbol{y}|\tilde{\boldsymbol{y}}; \boldsymbol{\theta}_r) = \prod_t P\left(\boldsymbol{y}_t|\boldsymbol{y}_{<t}, \tilde{\boldsymbol{y}}; \boldsymbol{\theta}_r\right), \quad (3)$$

where $\boldsymbol{y}_t$ is the t-th token of the sentence, and $\boldsymbol{\theta}_r$ is a set of parameters. The loss function of this task is as follows:

$$\mathcal{L}_{rec}(\boldsymbol{\theta}_r) = -\sum_{(\boldsymbol{y}, \tilde{\boldsymbol{y}}) \in \tilde{Y}} \log P(\boldsymbol{y}|\tilde{\boldsymbol{y}}; \boldsymbol{\theta}_r). \quad (4)$$

3.2.3 Language Modeling Task

This task facilitates zero-shot generation of $P(\boldsymbol{d}^{sim}|\mathrm{x})$ by jointly training the reconstruction decoder as a language model. Once the model captures correct complexity that guides the model to generate the desired simple texts, it's more likely for the model to ignore the wrongly shared complexity information. Similar to Eq. 3, we have:

$$P(\boldsymbol{y}|\boldsymbol{\theta}_l) = \prod_t P\left(\boldsymbol{y}_t|\boldsymbol{y}_{<t}; \boldsymbol{\theta}_l\right). \quad (5)$$

It is equivalent to masking the encoder out and ignoring the attention modules between the encoder and reconstruction decoder. The model is optimized by the following loss function:

$$\mathcal{L}_{lm}(\boldsymbol{\theta}_l) = -\sum_{\boldsymbol{y} \in Y} \log P(\boldsymbol{y}|\boldsymbol{\theta}_l). \quad (6)$$

Finally, we train the entire SimpDefiner by jointly minimizing the weighted sum of all above mentioned loss functions. And the overall loss function is calculated as:

$$\mathcal{L} = \lambda_\alpha \mathcal{L}_{gen} + \lambda_\beta \mathcal{L}_{rec} + \lambda_\gamma \mathcal{L}_{lm}, \quad (7)$$

where λ_α, λ_β, λ_γ are hyper-parameters.

3.3 Parameter-Sharing Scheme

For parameters in the decoders, we divided them into two parts, which are complexity-independent and complexity-dependent parameters. The former ones are shared between decoders, and the latter ones are not, as illustrated in Figure 3.

We now introduce the complexity-dependent layers, namely Complexity-Dependent Layer Normalization and Complexity-Dependent Query Projection.

Complexity-Dependent Layer Normalization Previous works (Dumoulin et al., 2017; Jin et al., 2020) demonstrated that the layer normalization is related to the style of the target texts. We further argue that as an attribute of style, the complexity can be retained by independent layer normalization. Thus, we make the scaling and shifting parameters

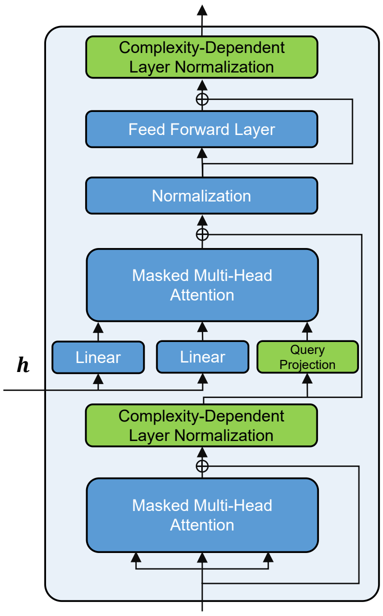

Complexity-Dependent Layer Normalization
Feed Forward Layer
Normalization
Masked Multi-Head Attention
Linear
Linear
Query Projection
h
Complexity-Dependent Layer Normalization
Masked Multi-Head Attention

Figure 3: The parameter-sharing scheme between decoders. Parameters in blue layers are shared, and those in green are not.

for layer normalization not shared in both decoders. This approach is to transform a layer activation $\boldsymbol{x}$ into a complexity-specific normalized activation $\boldsymbol{z}$ as:

$$\boldsymbol{z} = \gamma_c(\frac{\boldsymbol{x} - \mu}{\sigma}) - \beta_c, \tag{8}$$

where μ, σ are the mean and standard deviation of the batch of $\boldsymbol{x}$, and γ_c, β_c are learnable parameters specific to complexity c. Note that c is a binary variable indicating different decoders. This mechanism is used in all decoder layers.

Complexity-Dependent Query Projection The decoder layers extract information from encoded hidden states through cross-attention mechanism. We believe that the required information may vary for different complexity. Therefore, the parameters of the linear mapping used for the query transformation in the cross-attention are not shared among decoders. This calculation is as follows:

$$\boldsymbol{Q} = \hat{\boldsymbol{Q}} \cdot \boldsymbol{W}_c^q, \tag{9}$$

where $\boldsymbol{W}_c^q$ is the query transformation matrix specific to complexity c. The obtained query vector $\boldsymbol{Q}$ is then fed into the cross-attention mechanism. By using this approach, the model can obtain different information from the encoded hidden states for different complexities.

4 Datasets

We evaluate the proposed multitasking framework on both English and Chinese datasets. Each language has a definition generation dataset and a simple text corpus.

4.1 English Dataset

The English datasets are constructed from the Oxford Dictionary (OD) and Oxford Advanced Learner's Dictionary (OALD). Since the OALD is for language learners, it has much simpler definitions than OD. Therefore, we use the OD for the definition generation training, and use the OALD for validation of simple definition generation. Note that the words used for testing are excluded from the training and validation sets.

For the definition generation dataset, we directly use the OD dataset published by Gadetsky et al. (2018). The training set has 33,128 words and 97,855 entries. Each entry consists of a triplet of $(w^*, \boldsymbol{c}, \boldsymbol{d}^{com})$. For testing, we align the words and context in OD with the definitions in OALD through manual annotation. The annotated test set includes 3,881 words and 5,111 entries, which is used for automatic evaluation in experiments. Each entry in the test set has both golden complex and simple definitions from OD and OALD, respectively. Detailed statistics are listed in Table 1.

We extract the OALD definitions that are not in the test set for constructing the simple text corpus. This corpus has 32,395 sentences with an average length of 12.12. We list more statistics in Table 2.

During training, the definition generation dataset and the simple text corpus are randomly sampled as mini-batches respectively. And there is no alignment between the two mini-batches at each step.

4.2 Chinese Dataset

For the definition generation dataset, we use the Chinese WordNet (CWN) (Huang et al., 2010), which is a semantic lexicon aiming to provide a knowledge base of sense distinction.[2] We use the corresponding words, contexts, and definitions in CWN for the definition generation task. We split the entire dataset into training, validation, and test sets roughly according to the ratio of 8:1:1. The training set contains 6,574 words and 67,861 entries. Statistics are listed in Table 1.

[2] Chinese WordNet: http://lope.linguistics.ntu.edu.tw/cwn2

	OD			OALD	CWN		
	Train	Valid	Test	Test	Train	Valid	Test
Words	33,128	8,867	3,881	3,881	6,574	823	823
Entries	97,855	12,232	5,111	5,111	67,861	8,082	8,599
Context Length	17.74	17.80	16.24	16.24	34.49	34.73	34.06
Def. Length	11.02	10.99	10.03	12.74	14.76	14.60	14.72

Table 1: Statistics of the OD (English) dataset, OALD (English) test set, and CWN (Chinese) dataset. The rows are number of words and entries, and the average length of contexts and definitions.

	Sents	Tokens	Avg. Len
English	32,395	392,625	12.12
Chinese	58,867	860,761	14.62

Table 2: Statistics of simple text corpora. The columns are number of sentences and tokens, and the average length of sentences.

For the simple text corpus, we extract 58,867 sentences from a number of primary level Chinese as Second Language textbooks, with an average sentence length of 14.62.

Since no suitable dictionary can be used for evaluation, there are no golden simple definitions in Chinese Dataset. In the experiments, we count the difficulty level of words in definitions to estimate if they are simple. We also organize a manual evaluation to score the accuracy and simplicity of definitions.

5 Experiments

This section presents the experimental settings and evaluation methods.

5.1 Settings

Baselines We compare the SimpDefiner with generation-simplification pipelines. We first employ LOG-CaD (Ishiwatari et al., 2019) and MASS (Song et al., 2019) models to generate definitions, and then employ ACCESS (Martin et al., 2019) and MUSS (Martin et al., 2020) models to simplify them. Thus, we have four different pipeline baselines. Since these models are not available in Chinese, we only apply these pipelines to English datasets. For the Chinese SDG task, we specially pretrained a MASS-ZH model from scratch using the *Chinese Gigaword Fifth Edition*[3] corpus. Note that we set the learning rate to 3e-4, warmup steps to 500 when fine-tuning both MASS and MASS-ZH.

SimpDefiner We use the parameters in the MASS model to initialize the encoder and two decoders in SimpDefiner. For the sentence corruption in the text reconstruction task, we randomly delete or blank words with a uniform probability of 0.2, and randomly shuffle the order of words within 5 tokens. For the language modeling task, we set the input representations to **0** and use the simplified text as the target output. We tune the λ parameters in Eq. 7 on the validation set and adopt the same hyper-parameters as the baseline for comparison. We set 5 different random seeds as and report the average result of multiple runs. Each run takes 7.68 GPU hours on 4 GeFource RTX 2080 Ti GPUs.

5.2 Evaluation

Evaluation of the generated definitions mainly focuses on two aspects, i.e., accuracy and simplicity. We perform both automatic and manual evaluations for each aspect.

We first introduce these automatic metrics, and then the manual evaluation method.

BLEU Previous definition generation studies (Noraset et al., 2017; Yang et al., 2020; Kong et al., 2020) used the BLEU (Papineni et al., 2002) score to measure the closeness of generated results to the standard answers, and to evaluate the accuracy of results. Since the English test set is manually annotated, we calculate the BLEU score of both complex and simple definitions, respectively.

Semantic Similarity In addition to the BLEU score, we use the sentence-transformers toolkit (Reimers and Gurevych, 2020) to convert the generated definitions and references into sentence vectors, and calculate cosine similarity between them.

SARI SARI (Xu et al., 2016) is a lexical simplicity metric that measures how good are the words added, deleted and kept by a simplification model. This metric compares the model output to simplification references and the original sentence. We use

[3] https://catalog.ldc.upenn.edu/LDC2011T13

	Complex		Simple		SARI
	BLEU	SSim	BLEU	SSim	
LOG-CaD	19.04	40.32	–	–	–
+ ACCESS	–	–	12.32	32.63	38.02
+ MUSS	–	–	11.74	27.66	36.53
MASS	24.00	52.78	–	–	–
+ ACCESS	–	–	12.95	38.53	38.59
+ MUSS	–	–	12.58	37.49	38.48
SimpDefiner	**24.17**	**53.87**	**15.05**	**46.99**	**40.36**

Table 3: Main results on the English test set. LOG-CaD (Ishiwatari et al., 2019) is a definition generation model.

	L1-3 (%)	L7+ (%)
MASS	44.16	37.05
SimpDefiner	**48.03**	**36.59**

Table 4: Main results on the Chinese test set.

the SARI implementation in the EASSE toolkit[4].

HSK Level HSK, namely Chinese Proficiency Test, is set up to test the proficiency of non-native speakers[5]. It has nine levels, from easy to hard, and each level corresponds to a vocabulary. We count the proportion of words at levels 1-3 and 7+ in the generated definitions. The higher the proportion of words in levels 1-3 (7+), the easier (more challenging) the definitions are understood.

Manual Evaluation We randomly select 200 words and contexts from the Chinese test set and let the MASS and SimpDefiner generate definitions for them one by one. We mix the two generated definitions and the golden complex definition and then ask three native-speaker annotators to score them. Specifically, each annotator evaluates the definitions on two criteria of accuracy and simplicity. Both criteria have a range of 1-3. For accuracy, the annotators are asked to evaluate how semantically relevant the definitions are to the word. For simplicity, the annotators are asked to evaluate how simple the definitions are. After collecting evaluation results, we average the scores as final score.

6 Results and Analysis

6.1 Main Results

Table 3 and Table 4 present the experiment results on the English and Chinese test sets respectively. Results show that our proposed SimpDefiner significantly outperforms baseline methods of generation-simplification pipelines on both English and Chinese datasets.

		#1	#2	#3	Avg.
Acc.	Golden	3.00	2.93	2.98	2.97
	MASS	1.26	1.30	1.38	1.31
	SimpDefiner	1.48	1.47	1.59	**1.51**
Sim.	Golden	2.04	2.06	2.11	2.07
	MASS	1.92	2.03	1.89	1.95
	SimpDefiner	2.14	2.04	2.21	**2.13**

Table 5: Manual evaluation results on the Chinese test set. Accuracy and simplicity scores are listed in the table. The last column are averaged scores among all three annotators.

For English results, the performance of simple definition generation improves 2.1 and 8.46 on the BLEU and SemSim metrics respectively, and improves 1.77 on the SARI metric. This indicates that both accuracy and simplicity are effectively improved comparing with the baseline. We also observe that complex definition generation also improves by 0.17 on BLEU and 1.09 on SemSim. This shows that SimpDefiner improves the ability to generate both complex and simple definitions.

For Chinese results, we compute the HSK Level metric on generated simple definitions. We can see that the proportion of low-level (HSK level 1-3) words increases by 3.87%, and that of high-level (HSK level 7+) words decreases by 0.46%. The lexical complexity of the SimpDefiner generated definitions are significantly reduced.

Besides, we also conduct a manual evaluation on the Chinese test set, and the results are listed in Table 5. From the averaged scores, we observe that SimpDefiner outperforms MASS by 0.2 in terms of accuracy (more accurate) and 0.18 in terms of simplicity (more straightforward). On the accuracy score, all three annotators agree that SimpDefiner has higher accuracy than MASS, which shows the superiority of our framework. As expected, the golden definitions have the highest accuracy in the table, far exceeding the definitions generated by the two models. We believe this is caused by insufficient knowledge in the model, and this can be solved by using larger pretrained models, such as BART (Lewis et al., 2019). On the simplicity score, three annotators agree that SimpDefiner generates simpler definitions than MASS, and two of three annotators think SimpDefiner generates simpler definitions than the golden ones.

6.2 Ablation Study

We conduct ablation experiment to demonstrate the effectiveness of SimpDefiner components and

[4] `https://github.com/feralvam/easse`

[5] `http://www.chinesetest.cn`

ID	Model	Complex		Simple		SARI
		BLEU	SSim	BLEU	SSim	
①	SimpDefiner	24.31	53.60	**15.24**	**47.05**	**40.19**
②	① – LM	23.83	53.04	14.82	45.74	39.63
③	② – TR	**25.02**	53.80	13.66	44.01	38.58
④	① – LN	24.45	53.76	13.87	44.66	38.61
⑤	① – QP	23.40	52.95	14.61	45.57	39.87
⑥	④ – QP	24.80	**53.95**	13.90	44.77	38.52

Table 6: Ablation study on the English test set. LM: the language modeling task. TR: the text reconstruction task. LN: complexity-dependent layer normalization. QP: complexity-dependent query projection.

$(\lambda_\alpha, \lambda_\beta, \lambda_\gamma)$	Complex		Simple		SARI
	BLEU	SSim	BLEU	SSim	
(0.8,0.1,0.1)	**24.31**	**53.60**	**15.24**	**47.05**	40.19
(0.6,0.2,0.2)	23.27	53.19	15.01	46.85	40.49
(0.4,0.3,0.3)	21.92	51.82	15.11	46.54	**40.74**

Table 7: Different hyper-parameter settings on the English test set.

Word	commander
Context	Military commanders have warned coalition troops in the south.
Golden	A person who is in charge of sth, especially an officer in charge of a particular group of soldiers or a military operation.
Baseline	An officer of the highest rank is a country in a country.
SimpDefiner	The head of a military force.
Word	督促 (supervise and urge)
Context	我很感谢他的支持、鼓励与督促。 I appreciate his support, encouragement and supervision.
Golden	监督他人并促使后述事件发生。 Supervise others and promote the occurrence of the following events.
Baseline	以后述对象为凭借进行特定事件。 Sb. is used as a reference for specific events.
SimpDefiner	要求后述对象赶快行动。 Ask sb. to act quickly.

Table 8: Cases of generated simple definitions.

the parameter sharing scheme. For the language modeling (LM) and text reconstruction (TR) tasks, we ablate them by setting their weights to 0. For the layer normalization (LN) and query projection (QP) as parameter-shared layers, we ablate them by sharing their parameters between models. We illustrate the experiment results in Table 6.

In general, ablating any of the components or parameter-shared layers reduces the performance in terms of simple definitions, which indicates that the SimpDefiner benefits from both components and parameter sharing scheme. We also observe that the performance of ablation experiments have slight disturbance on complex definitions. But since we pay more attention to the performance on simple definitions, we argue that the benefits of SimpDefiner far outweigh the losses.

6.3 Analysis on Hyper-Parameters

Furthermore, we conduct additional experiments on the English dataset to study how hyper-parameters affect the performance. By setting different λ to each model, we observe the relationship between the performance and these weights.

The experiment results are listed in Table 7. From the table, we observe the inconsistency between metrics. As the definition generation task weight declines, the BLEU and SemSim metrics are generally declining, but the SARI metric is increasing. Since the BLEU and SemSim measure the accuracy and the SARI measures simplicity, we consider this phenomenon as a seesaw between the two attributes of accuracy and simplicity. The balance between them can be achieved by conditioning the hyper-parameters.

6.4 Case Study

Table 8 shows two generation cases from English and Chinese test set respectively. In both cases, the golden definition is a long sentence with quite complicated syntax. The baseline generated definitions contains difficult words and often wrongly defines the given word. In the English case, the word *commander* is defined by the baseline as *an officer of the highest rank in a country*, which is incorrect in most cases. In the Chinese case, the baseline generated definition contains difficult words like 凭借 *(reference)* and 特定事件 *(specific events)*. On the other hand, the SimpDefiner generates simple and accurate definitions in both cases.

7 Conclusion

In this work, we propose the SDG task, a novel task of generating simplified definitions in a zero-shot manner. To this end, we leverage a multitasking framework SimpDefiner to tackle this task. We introduce a text reconstruction task to the framework to control the text complexity, and a language modeling task to enhance the decoder. For evaluation, we construct a novel test set in English by manually aligning the two dictionaries of OD and OALD. The automatic and manual evaluations indicate that the our proposed framework can generate more accurate and more straightforward definitions than other models and the generation-simplification pipelines. In the future, we will try to combine the current method with prompt learning methods, aiming to let users condition the complexity of generated definitions.

Acknowledgements

This work was supported by the funds of Beijing Advanced Innovation Center for Language Resources (No. TYZ19005), Research Project of the National Language Commission (No. ZDI135-131) and National Natural Science Foundation of China (No. 62106138, No. 61872402). We would like to thank Xiaowan Wang, Chenhui Xie, and Junhui Zhu for their manual evaluation and all anonymous reviewers for their valuable comments and suggestions on this work.

References

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2016. Breaking sticks and ambiguities with adaptive skip-gram. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Ting-Yun Chang, Ta-Chung Chi, Shang-Chi Tsai, and Yun-Nung Chen. 2018. xsense: Learning sense-separated sparse representations and textual definitions for explainable word sense networks. *CoRR*, abs/1809.03348.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. 2017. A learned representation for artistic style. In *5th International Conference on Learning Representations, ICLR 2017*.

Fatemeh Enayati and Abbas Pourhosein Gilakjani. 2020. The impact of computer assisted language learning (CALL) on improving intermediate EFL learners' vocabulary learning. *International Journal of Language Education*, 4(2).

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Dynamic multi-level multi-task learning for sentence simplification. *CoRR*, abs/1806.07304.

Chu-Ren Huang, S. Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I. Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese wordnet : design, implementation, and application of an infrastructure for cross-lingual knowledge processing. *Journal of Chinese information processing*, 24.

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. Definition modelling for appropriate specificity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. Learning to describe unknown phrases with local and global contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

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*.

Takashi Kawashima and Tomohiro Takagi. 2019. Sentence simplification from non-parallel corpus with adversarial learning. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE.

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2).

Cunliang Kong, Liner Yang, Tianzuo Zhang, Qinan Fan, Zhenghao Liu, Yun Chen, and Erhong Yang. 2020. Toward cross-lingual definition generation for language learners. *CoRR*, abs/2010.05533.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Jiahuan Li, Yu Bao, Shujian Huang, Xinyu Dai, and Jiajun Chen. 2020. Explicit semantic decomposition for definition generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Yuri Lolita, Endry Boeriswati, and Ninuk Lustyantie. 2020. The impact of computer assisted language learning (CALL) use of english vocabulary enhancement. *Linguistic, English Education and Art Journal*, 4(1).

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2020. Multilingual unsupervised sentence simplification. *CoRR*, abs/2005.00352.

Louis Martin, Benoît Sagot, Éric de la Clergerie, and Antoine Bordes. 2019. Controllable sentence simplification. *CoRR*, abs/1910.02677.

Timothee Mickus, Denis Paperno, and Matthieu Constant. 2019. Mark my word: A sequence-to-sequence approach to definition modeling. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, Turku, Finland. Linköping University Electronic Press.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2020. VCDM: Leveraging Variational Bi-encoding and Deep Contextualized Word Representations for Improved Definition Modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Eleanor Rosch and Carolyn B Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4).

Thomas M Segler, Helen Pain, and Antonella Sorace. 2002. Second language vocabulary acquisition and learning strategies in ICALL environments. *Computer Assisted Language Learning*, 15(4).

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*. PMLR.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*.

Andrea Tyler and Vyvyan Evans. 2001. Reconsidering prepositional polysemy networks: The case of over. *Language*.

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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Tu Vu, Baotian Hu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Sentence simplification with memory-augmented neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4.

Liner Yang, Cunliang Kong, Yun Chen, Yang Liu, Qinan Fan, and Erhong Yang. 2020. Incorporating sememes into chinese definition modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Yihua Zhang. 2011. Discussion on the Definitions in Chinese Learner's Dictionaries: Comparative Study of Domestic and Foreign Learner Dictionaries (Translated from Chinese). *Chinese Teaching in the World*.