Fine-grained Contrastive Learning for Definition Generation

Hengyuan Zhang[1*], Dawei Li[2*], Shiping Yang[3], Yanran Li[4†]

[1]Shenzhen International Graduate School, Tsinghua University
[2]Halicioğlu Data Science Institute, University of California, San Diego
[3]School of Computer Science, Beijing University of Posts and Telecommunications
[4]Independent Researcher
zhang-hy22@mails.tsinghua.edu.cn, dal034@ucsd.edu,
yangshiping@bupt.edu.cn, yanranli.summer@gmail.com

Abstract

Recently, pre-trained transformer-based models have achieved great success in the task of definition generation (DG). However, previous encoder-decoder models lack effective representation learning to contain full semantic components of the given word, which leads to generating under-specific definitions. To address this problem, we propose a novel contrastive learning method, encouraging the model to capture more detailed semantic representations from the definition sequence encoding. According to both automatic and manual evaluation, the experimental results on three mainstream benchmarks demonstrate that the proposed method could generate more specific and high-quality definitions compared with several state-of-the-art models.

1 Introduction

When readers find some expressions unfamiliar during reading a text, machines can help. The task of Definition Generation (DG) aims to generate a textual definition for a given word or phrase (the target), according to a surrounding context (the local context) (Ni and Wang, 2017). In addition to assisting readers in comprehending expressions, the task of DG is also useful for generating definition when building dictionaries.

Recently, pre-trained encoder-decoder models have achieved great successes on this task (Huang et al., 2021; Kong et al., 2022). Despite their successes, the definitions produced by these pre-trained models often contain several types of errors (Noraset et al., 2017; Huang et al., 2021). According to Table 1, "under-specific problem" is the most frequent error that the generated definition conforms to the general semantics but loses certain parts of meaning of the target word. As presented in Table 2, the definition produced by T5 model is

Error Types	Ratio
Under-spcified	**9.0%**
Over-specified	5.5%
Self-reference	3.0%
Wrong part-of-speech	1.0%
Opposite	1.0%

Table 1: Ratio of each error type of the definitions generated in Huang et al. (2021).

word	double
Reference	twice as great or many
Generated	characterized by two equal parts or components

Table 2: The definition of word "double", where *Reference* is from WordNet dictionary and *Generated* is by T5-Base of Huang et al. (2021).

under-specific as it omits the meaning of *great* in the word "double" under the context "*ate a double portion*". The under-specific problem harms the accuracy of the generated definitions and in turn limits the applications of definition generation techniques in many scenarios.

This problem is partially attributed to the decoder's inability to fully extract the semantic components from the word encoding (Li et al., 2020a). For pre-trained encoder-decoder models, they focus on restoring and denoising the whole text in the pre-training stage, rather than learning fine-grained semantic representation of a single word or phrase (Lewis et al., 2019; Bi et al., 2020; Shao et al., 2021). In other words, the pre-trained encoder-decoder models are ineffective in capturing rich semantic components for the given word thus leading to generating under-specific definitions.

To remedy the under-specific problem in pre-trained definition generation models, we get inspired from contrastive learning method (Radford

---

et al., 2021; Li et al., 2020b) and propose a novel definition generation method based on a designed contrastive objective. Conceptually, definition generation is to transform the encoding of the target word to its textual interpretation. To this end, the encoding and the decoding of the target word can be regarded as two views of representations with respect to the same semantics. Our idea is then to leverage the two representations in the definition generation model, and encourage them to align with each other to capture fine-grained semantics. Specifically, we treat the target word representation and the definition representation as a positive pair, and feed them into a contrastive learning objective. This kind of contrastive loss is naturally complementary for the language generation loss, and can be seamlessly incorporated into existing pre-trained encoder-decoder models.

To validate the effectiveness of our proposal, we conduct a series of experiments on three publicly available datasets. Both automatic and manual evaluation results suggest that our method generates more specific definitions and addresses well the under-specific problem in the task of definition generation. In general, our contributions can be summarized as follows:

- We tackle the under-specific problem for pre-trained definition generation models by developing a novel fine-grained contrastive learning objective.

- We validate the effectiveness of the proposed method through comparing with several SOTA models on three popular datasets using both automatic and manual judgments.[1]

- We analyze the details of our method by performing ablated studies and demonstrate the effect of our method in addressing under-specific problem based on case studies.

2 Related Work

2.1 Definition Generation

The task of Definition Generation is firstly proposed by Noraset et al. (2017). They used word embedding to generate its corresponding definition, and utilize definition generation as an auxiliary task for reverse dictionary and word embedding training.

---

[1] Our code could be found in https://github.com/rattlesnakey/Definition-Gneration-Contrastive

Some later works explore more application scenarios and model architectures for definition generation. Ni and Wang (2017) propose a dual-encoder model to generate the proper definition of the given word under a specific context, and use it for explaining emerging words on the Internet. Gadetsky et al. (2018) use both local and global information of the words in their model for word disambiguation. Following them, Ishiwatari et al. (2019) design gate mechanisms to fuse multi-source information of the word and context. Furthermore, some works attempt to utilize other information of the target word. Washio et al. (2019) build relation of defined and defining words using word pair embedding (Joshi et al., 2018). Different from former works that using distributed representations of target words, Yang et al. (2019) introduce target words' concepts in HowNet (Dong and Dong, 2003) as fine-grained knowledge in Chinese definition modeling. Also, there exist literature works based on refined methods to learn the target words. Both Li et al. (2020a) and Reid et al. (2020) decompose the meaning of the target word into a group of latent variables and rely on variational inference for estimation.

Recently, pre-trained encoder-decoder models have been used in definition generation and achieved great success. Bevilacqua et al. (2020) use special tokens to mark the target word in the context and feed them into a BART model (Lewis et al., 2019). Huang et al. (2021) fine-tune a T5 model and re-rank all the candidate results from the T5 model to obtain definitions in a proper specificity. Kong et al. (2022) design a MASS model based on multi-task framework to generate simple definition in an unsupervised manner. Despite of their promising performances on definition generation, the under-specific problem has been less investigated. Although Huang et al. (2021) design a scoring mechanism that measures definitions' specificity, we argue that the fundamental reason of the under-specific problem lies in the lack of fine-grained semantic learning in pre-trained encoder-decoder models, which we leverage contrastive learning to address in this work.

2.2 Contrastive Learning in Semantic Representation

Contrastive learning has been widely used in enhancing semantic information for various NLP tasks. For example, Gao et al. (2021) use a dropout trick to derive positive samples in the embedding

level, and then apply both supervised and self-supervised methods to acquire better sentence embedding. Radford et al. (2021) use contrastive learning to pre-train a vision language model to align the message between images and their corresponding text. Li et al. (2022) use masked language modeling and contrastive learning to perform multi-task pre-training, and demonstrate that contrastive learning benefits in connecting word gloss and its corresponding vectors. Li et al. (2020b) and Srivastava and Vemulapati (2022) implement contrastive learning as an auxiliary task to encourage the transformer encoder better capture the semantic alignment.

In this work, we borrow the idea of using contrastive methods in semantic representation learning. For a given target word, there are two representations in the task of definition generation: the word representation generated by the encoder, and the definition representation produced by the decoder. These two kinds of representations can be regarded as two views of the semantics of the target word to be explained. By aligning the representation spaces between the encoder and the decoder using contrastive learning, we force the model to pay much attention to the fine-grained semantic information during representation learning. In this way, the under-specific problem will be mitigated when using pre-trained encoder-decoder models to generate definitions.

3 Method

In this section, we present our method of using contrastive learning to enhance target words' representation for definition generation. Specifically, we first formulate the definition generation task and introduce the denotations (Section 3.1). Then we provide a preliminary description of the definition generation processing based on T5 (Section 3.2). Finally, we explain how to apply the contrastive loss in the training process to solve the under-specific problem and improve the generation quality (Section 3.3). Figure 1 depicts the overview pipeline of our method.

3.1 Task Formulation

Given a word or phrase $W = \{w_i, ..., w_j\}$ and its surrounding context $C = \{w_0, ..., W_k\}(0 < i < j < k)$, the task of definition generation is to generate the definition $D = \{d_0, ...d_T\}$ to explain the meaning of W under C. This process can be formulated as:

$$P(D|W, C) = \prod_{t=0}^{T} p(d_t|d_{<t}, W, C) \quad (1)$$

3.2 Definition Generation with T5

Our work aims at addressing the under-specific problem when using pre-trained encoder-decoder models for definition generation. Without loss of generality, we take T5 (Raffel et al., 2020) as our backbone model, which is a transformer-based encoder-decoder model trained on large-scale corpus, and has demonstrated its effectiveness on definition generation task (Huang et al., 2021).

To apply T5 for definition generation, we first concatenate the target word and the given context together behind the prefix prompts "word:" and "context:" respectively. The concatenated input is then fed to the T5 encoder with L_E layers of encoder block E_Block. Then we get the last hidden state $\mathbf{H}^{L_E}$, which contains the semantic information of the target word and local context:

$$\mathbf{H}_0 = \mathrm{Emb}(\mathrm{Splice}(W, C)) \quad (2)$$

$$\mathbf{H}^l = \mathrm{E\_Block}(\mathbf{H}^{l-1}), l \in [1, L_E] \quad (3)$$

Here W stands for the target word, C for the given context, and Splice is the operation to concatenate the target word and the given context with their corresponding prefixes. Also, Emb is the Embedding layer that converts the input tokens into embedding vectors.

After encoding, the T5 decoder will learn to generate an appropriate definition conditioned on encoding $\mathbf{H}^{L_E}$ and the previous generation result. During decoding, the teacher-forcing mechanism is applied to guarantee the previous information being attended at the current step t:

$$\mathbf{G}_t^0 = \mathrm{Emb}(D_t) \quad (4)$$

$$\mathbf{G}_t^l = \mathrm{D\_Block}(\mathbf{H}^{L_E}, \mathbf{G}_{\leq t}^{l-1}), l \in [1, L_D] \quad (5)$$

Here D_t represents the tth token in the definition sequence. After passing through L_D layers of the decoder block D_Block, we get the decoder's last hidden state $\mathbf{G}^{L_D}$.

Finally, a softmax function is added upon a linear head to transform $\mathbf{G}^{L_D}$ into a prediction distribution matrix $\mathbf{V} \in \mathbb{R}^{|V| \times |D|}$. Here $|V|$ and $|D|$
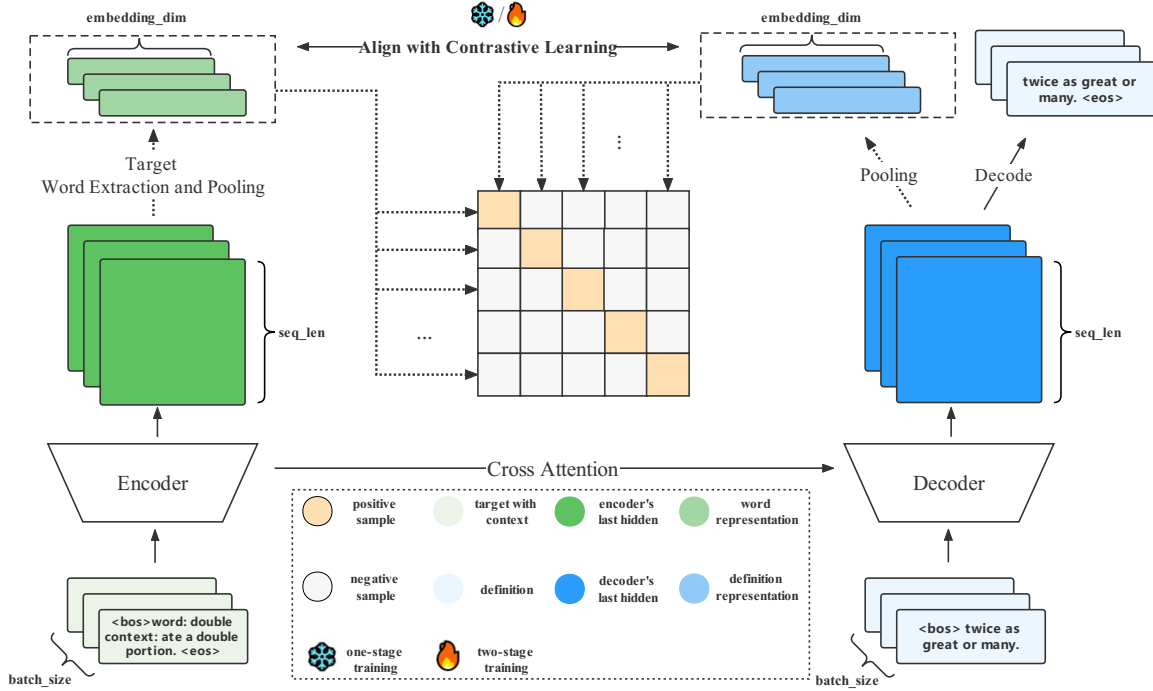
Figure 1: The overview training process of our proposed model. The solid arrows indicate the data-flow of maximum likelihood estimate learning, and the dash arrows indicate the data-flow of contrastive learning. Note that the snow icon represents the one-stage training where the model is trained from scratch with the contrastive and generation loss. The fire icon represents the two-stage training, where at the first stage the model is fine-tuned only by the generation loss, and at the second stage the contrastive and generation loss are together applied then.

stand for the vocabulary size and the length of the ground-truth definition, respectively. To optimize, a cross-entropy loss is applied to measure the discrepancy between the generated distribution and the ground-truth distribution.

3.3 Fine-grained Modeling with Contrastive Learning

Here we describe our proposal of applying contrastive learning for definition generation. Conceptually, definition generation requires the model to understand the target word and produce its definition to explain the meaning of the word in the context. To this end, definition generation can be cast as a mapping between the understanding of the target word (in the encoder side) and the generation of the word definition (in the decoder side).

Hence, our idea is to leverage the representations obtained from both the encoder and decoder side in the model, and encourage them to align with each other to capture fine-grained semantics. By regarding the representations from both sides as two views of the target words' semantics, we are able to deploy a contrastive loss during the generation process in the training phase.

Formally, we denote the target word encoding generated by the encoder in T5 as $\mathbf{H}_{target}$, and the definition encoding generated by the decoder as $\mathbf{G}^{L_D}$. In general, target word encoding $\mathbf{H}_{target}$ is obtained by extracting the encoding of the target word's position in $\mathbf{H}^{L_E}$, and definition encoding $\mathbf{G}^{L_D}$ is generated by the decoder to decode and get the definition sequence later.

After encoding, we use a pooling function f() to aggregate the $\mathbf{H}_{target}$ and $\mathbf{G}^{L_D}$ respectively, and obtain target word representation $\mathbf{h}$ and definition representation $\mathbf{g}$ with the same length:

$$\mathbf{h} = \mathrm{f}(\mathbf{H}_{target}) \tag{6}$$

$$\mathbf{g} = \mathrm{f}(\mathbf{G}^{L_D}) \tag{7}$$

Note that there are multiple choices to implement the pooling function f(). Empirically motivated, we adopt max-pooling $\mathrm{Max}()$ and achieve our best performance in the main experimental results. We also present the results with mean-pooling $\mathrm{Mean}()$ in the ablation study in the following sections.

Eventually, we treat the two representations $\mathbf{h}$ and $\mathbf{g}$ in the same sample as a positive pair, and

1004

	WordNet			Oxford			Urban		
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
Phrases	7,938	998	1,001	33,128	8,867	8,850	190,696	26,876	25,797
Entries	13,883	1,752	1,775	97,855	12,232	12,232	411,384	57,883	36,450
Context length	5.81	5.64	5.77	17.74	17.80	17.56	10.89	10.86	11.22
Desc. length	6.61	6.61	6.85	11.02	10.99	10.95	10.99	10.95	12.05

Table 3: Statistics of The Datasets.

define our contrastive learning stage's training objective as follows:

$$L_C = \sum_{i=1}^{N} -log \frac{e^{\text{sim}(h_i,g_i)}/\tau}{\sum_{j=1}^{N} e^{\text{sim}(h_i,g_j)}/\tau} \quad (8)$$

where N denotes a mini-batch of training samples. The τ is a temperature hyper-parameter and $\text{sim}(,)$ stands for the cosine similarity function. During learning, the contrastive loss in Eq. 8 enforces the model to concentrate on the discrepancy between the two views of the same semantic unit, i.e., the target word.

3.4 Two-Stage Training

In addition to the newly introduced contrastive loss, we also train the model based on the commonly adopted generation loss, which takes advantage of language modeling ability.

As depicted in Figure 1, our full training strategy follows a two-stage paradigm. At the first stage, we finetune our model only with the generation loss. In the second stage, we combine the contrastive loss in the training and optimize the model with mixed loss L_{Final}:

$$L_{Final} = \lambda * L_C + (1 - \lambda) * L_G \quad (9)$$

where λ is a hyper-parameter to balance the two loss. The two-stage training allows to incrementally train the decoder learn the semantic information from the definition sequence at the very beginning, and guarantees the quality of the definition encoding for the encoder to discriminate in the following stage.

By combining the contrastive loss with generation loss, our method is able to: (1) learn fine-grained representation for the target word, (2) mitigate the under-specific problem in the encoder-decoder models, and (3) improve the overall quality of the generated definition.

4 Experiments

In this section, we compare our method with several state-of-the-art methods and conduct a series of experiments to verify the effectiveness of our method in addressing the under-specific problem in definition generation.

4.1 Datasets

For evaluation, we follow previous works and acquire three popular datasets, which are ensembled by Ishiwatari et al. (2019)[2]. Each entry in a dataset consists of three elements: (1) a target word or phrase, (2) the corresponding definition, and (3) one usage example of the target as a local context. If a target has multiple definitions and examples, we treat them as different entries. For fair comparison, each dataset is split into *train*, *dev* and *test* sets according to Ishiwatari et al. (2019). The statistics of these datasets are shown in Table 3.

WordNet dataset The Wordnet dataset is collected by Noraset et al. (2017) from the Wordnet dictionary and the GNU Collaborative International Dictionary of English[3]. In this work, we follow Ishiwatari et al. (2019) and use the extended version of WordNet dataset, where usage examples for each entry are added and the entries without usage examples are removed.

Oxford dataset The Oxford dataset is collected using APIs of Oxford Dictionaries[4] by Gadetsky et al. (2018).

Urban dataset The Urban dataset is collected from Urban Dictionary[5], which is the largest online slang dictionary. Unlike the former two datasets, this dataset contains many non-standard phrases

[2] http://www.tkl.iis.u-tokyo.ac.jp/~ishiwatari/naacl_data.zip
[3] http://wwwgcide.gnu.org.ua
[4] https://developer.oxforddictionaries.com
[5] https://www.urbandictionary.com

1005

	WordNet		Oxford		Urban	
	BLEU	NIST	BLEU	NIST	BLEU	NIST
I-Attention	23.77	44.30	17.45	35.79	8.81	19.43
Local	24.78	40.32	17.58	31.30	8.99	17.39
Global	23.59	49.70	14.95	32.79	5.15	10.45
LOG-CaD	25.19	43.54	18.57	38.22	9.93	19.29
T5-Reranking	**32.72**	64.57	26.52	74.17	17.71	35.53
T5-Contrast (Ours)	32.05$_{(-2.1\%)}$	**74.71**$_{(+15.5\%)}$	**27.11**$_{(+2.2\%)}$	**79.42**$_{(+7.1\%)}$	**19.44**$_{(+9.8\%)}$	**41.01**$_{(+15.4\%)}$
T5-Base	31.72	57.35	25.44	66.92	17.66	26.86

Table 4: Automatic evaluation results on test sets of three datasets. The best results in each dataset are in bold. We also add the quantitative comparison results between our method and the strongest baseline model T5-Reranking.

with more than one word. In Urban dataset, all terms, definitions, and examples are submitted by users on the Internet.

4.2 Compared Models

To evaluate the effectiveness of our method, we compare with the following models:

Global (Noraset et al., 2017) is the first definition generation technique that only accesses the global context of the target word.

Local (Ni and Wang, 2017) is the refined model that utilizes both word-level and character-level information to get the target word encoding based on the surrounding context.

I-Attention (Gadetsky et al., 2018) combines local and global contexts together and employs latent variable modeling and soft attention mechanisms.

LOG-CaD (Ishiwatari et al., 2019) integrates the designs in the previous methods and uses gate-mechanism to balance information from different sources in the decoding phase.

T5-Reranking (Huang et al., 2021) is the current SOTA method in definition generation. It uses a pre-trained T5 to get generation results first and designs a score mechanism to measure and sample definitions in appropriate specificity.

T5-Base Besides, we also fine-tune a pre-trained T5 only using the generation loss we mention in Section 3.4 as a baseline (denoted as T5-Base).

4.3 Automatic Metrics

Following common practice, we adopt two automatic evaluation metrics to assess the quality of the definitions generated by each model.

BLEU The metric BLEU (Papineni et al., 2002) has been widely used in previous works to measure the closeness between the generated results and human reference. It measures the geometric average

of the precision over hypothesis n-grams with an additional penalty to discourage short definition.

NIST NIST (Doddington, 2002) is similar to BLEU, but considers up-weighting rare, informative n-grams. We use NLTK[6] tool to calculate NIST metric.

4.4 Experimental Setups

We train all models in PyTorch[7] (Paszke et al., 2019), and use the HuggingFace[8] (Wolf et al., 2019) implementation of T5. We train each model on a V100 GPU. For compared models, we replicate experiments following the implementations details released by Huang et al. (2021). For training our model, we use the base version of T5 with the same size of Huang et al. (2021). For each dataset, we finetune it using Adam (Kingma and Ba, 2014) optimizer with an initial learning rate of 3e-4 and the batch size of 16. In all the experiments, we train our model with a two-stage strategy as described in the previous section. Please refer to Appendix A for the detailed training settings in each stage, like max-epoch and early-stop threshold.

4.5 Main Results

Table 4 shows the automatic comparison results of each compared model on the three datasets. Considering the absolute scores, the proposed method T5-Contrast significantly outperforms other 5 models on almost every metric across the three datasets. Although the BLEU score on WordNet dataset obtained by our method is slightly lower (2.09%) than T5-Reranking (Huang et al., 2021), the NIST score of our method in WordNet dataset is notably higher (15.70%) than theirs. This strongly demonstrates the effectiveness and generalization of the proposed

---

[6]https://www.nltk.org
[7]https://github.com/pytorch/pytorch
[8]https://github.com/huggingface/

	WordNet		Oxford		Urban	
	BLEU	NIST	BLEU	NIST	BLEU	NIST
Ours	**32.05**	**74.71**	27.10	79.42	**19.44**	**41.01**
w/ $\mathrm{Mean}()$	31.07	71.48	**27.13**	**80.33**	18.57	40.29
w/ One-stage training	31.75	73.79	27.06	79.90	16.49	31.46
T5-Base	31.72	57.35	25.44	66.92	17.66	26.86

Table 5: Ablation study results on test sets. The best numbers are in bold.

method in generating high-quality definitions for a given word under a context.

It is obvious that the two refined T5 model, e.g., T5-Reranking (Huang et al., 2021) and T5-Contrast (ours) are the best and the second best model. By comparing the relative increases between these two models, we notice that our method T5-Contrast improves a lot on Urban dataset (9.8% relative increase on BLEU, and 15.4% relative increase on NIST). As compared to the datasets WordNet and Oxford, Urban dataset is more challenging due to the targets in it are often phrases, and the definitions are often long and complex. Drawing on the great promotion by T5-Contrast (ours) on the difficult dataset, we highlight the necessity of modeling fine-grained semantic in pre-trained models for definition generation.

4.6 Ablation Study

As introduced in Section 3, there are two novel designs in our method: (1) the contrastive learning with a pooling function, and (2) a two-stage training strategy that combines both generation loss and the contrastive loss. In this subsection, we conduct an ablation study to examine the variants of each component in the proposed method.

As shown in Table 5, replacing the pooling function $Max()$ with the mean-pooling $Mean()$ will bring in different changes on different datasets. Whereas the automatic scores drop a lot on WordNet and Urban datasets, they increase a bit on Oxford dataset. This indicates that the choice of pooling function might be empirically motivated, and in general the effect of contrastive learning does not vary a lot when the pooling function changes.

Moreover, we also examine the importance of two-stage training by removing the first stage of generation-only training and directly training our model using the combined loss (One-stage training). Especially on the challenging Urban dataset, the performance dramatically decreases when train-

ing T5 from scratch using the combined loss. Last but not least, each of our ablated variant still surpasses T5-Base on most metrics, which indicates the method's robustness.

4.7 Analysis on Hyper-Parameter

To explore how our method would be affected by the choice of the hyper-parameter λ in Eq. 9, we remain other settings the same as we mentioned in Section 4.4 and set different λ for each model to observe the performance change. The results on the Oxford dataset are reported in Table 6. As shown, when λ is set to 0.0, the model is "degraded" to the compared T5-Base model. Considering T5-Base model is fine-tuned only using the generation loss in our setting, it is identical to a variant without contrastive loss in the second training stage. To this end, their performances are the same. Also, the performance of the model when λ is set to 1.0 (without generation loss in the second training stage) is pretty bad. We attribute it to the fact that our task requires the ability of language generation and thus still need generation loss to guide contrastive learning in the right way. Besides the above extreme values of λ, we find the model achieves better performance when λ is higher (λ=0.8 and λ=0.6). It further illustrates that after the first stage of generation-only training, the model will benefit more from our fine-grained contrastive learning.

We also investigate the influence of training batch size on our method. We set our training batch size $\in \{8, 16, 32, 64\}$ and conduct experiments on the Oxford dataset. As Figure 2 shows, each model's performance in different batch size settings doesn't show much difference. It is probably due to our proposed method base on pre-trained T5 which has good prior knowledge.

4.8 Manual Evaluation

To adequately evaluate the generated definitions, we also adopt three kinds of manual metrics: (1)

λ	BLEU	NIST
1.0	7.71	25.67
0.8	27.11	79.42
0.6	27.23	79.86
0.4	26.66	77.68
0.2	26.54	78.60
0.0	25.44	66.92

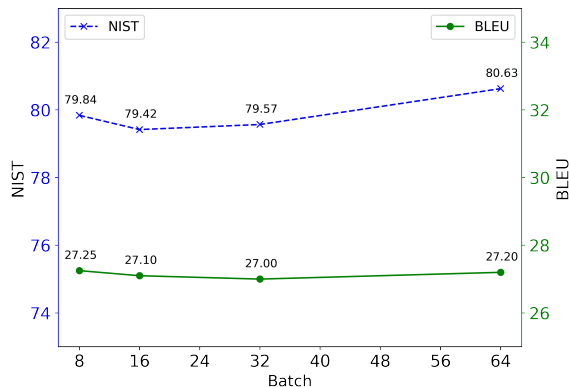Table 6: Different λ settings on Oxford test set.

Figure 2: Training batch size analysis results on Oxford test set.

Acccuracy (Acc.) (Ishiwatari et al., 2019) learning measures the semantic similarity between the generated definitions and the target words; (2) Fluency (Flu.) evaluates the readability of the generated definitions without considering the semantic alignment; (3) Under-specified (Under-spec.) (Noraset et al., 2017) calculates the ratio of under-specific definitions in the generated cases, which is curated to assess the model's capabilities in addressing the under-specific problem. The lower the ratio is, the better the model is in capturing fine-grained semantics during definition generation. Note that both Acc. and Flu. metrics are likert-scale of {0,1,2,3,4,5}.

Considering the labor resource cost, we conduct manual evaluation on Oxford dataset, and only compare our method with the strongest baseline T5-Reranking and the backbone T5-Base. For fairness, we randomly select 100 samples, acquire the generation results of each compared model, and pair them with the Golden definitions. Then we ask three well-trained annotators with at least CET (College English Test) 6-level English skills to rate the generated definitions according to the three manual metrics. At last, each model's score is the average of the three annotators' rates and the agreement among the annotators is ICC 0.962 with

(p<0.001) (Bartko, 1966), which indicates the results are reliable enough.

According to Table 7, the definitions generated by the proposed method T5-Contrast are better than those by other two models in terms of all the three metrics. Notably, the under-specific ratio significantly drops from 7.6% (T5-Base) to 4.8% (Ours). The manual evaluation results imply that the definitions produced by our method are more accurate, fluent, and fine-grained as compared to other pre-trained models.

	Acc.	Flu.	Under-spec.
T5-Base	3.17	3.89	7.6%
T5-Reranking	3.43	3.95	5.4%
T5-Contrast (Ours)	3.46	4.03	4.6%
Golden	4.57	4.92	0.2%

Table 7: Manual evaluation results on Oxford dataset.

4.9 Case Study

For better understanding, we show some example definitions generated by these compared models in Table 8. It is obvious that T5-Base produces an under-specific definition "*a positive criticism*" for the target word "*praise*" in the context *he always appreciated praise for this work*. The generated definition roughly expresses the positive meaning of the target word *appreciate*, but fails to provide the accurate meaning of *approval and commendation* in *praise*. In this case, this example definition by T5-Base is under-specific. As for T5-Reranking, it generates the word "*goodwill*", which is a multi-sense word where the one sense is "a kindly feeling of support" and the other sense is "the favor or advantage of a business". As such, this definition by T5-Reranking is also inaccurate to describe the word *praise*. On the contrary, the definition generated by our model that "*an expression of admiration or approval*" is more specific, which shows the effectiveness of our proposed method to remedy the under-specific problem. Due to the space limit, we give more sampled examples in Appendix B.

It is also worth noting that, with our contrastive learning loss, some definitions generated in the test time are even identical with their ground truths. This also supports our idea that fine-grained contrastive learning will benefit the pre-trained encoder-decoder models in modeling and generating definitions. We also put these kinds of cases

in Appendix C.

Word	Praise
Context	He always appreciated **praise** for his work.
T5-Base	A positive critisism.
T5-Reranking	An act to express goodwill.
Ours	An expression of admiration or approval.
Ground Truth	An expression of approval and commendation.

Table 8: An example showing the two generated definitions for the word "praise" by our model, T5-Base and T5-Reranking. The green text represents the appropriate specificity of the generated definition, and the text in red represents the hints where the generated definition is under-specific.

5 Conclusion

In this work, we tackle the under-specific problem when using pre-trained encoder-decoder models for definition generation. To address, We propose a fine-grained contrastive method to inject detailed semantic information into the model. Through extensive experiments, we demonstrate the effectiveness and generalization of the proposed method using both automatic and manual evaluations on three datasets.

In the future, we aim to introduce more fine-grained methods and language resources into definition generation.

References

John J Bartko. 1966. The intraclass correlation coefficient as a measure of reliability. *Psychological reports*, 19(1):3–11.

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 (EMNLP)*, pages 7207–7221.

Bin Bi, Chenliang Li, Chen Wu, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2020. Palm: Pre-training an autoencoding&autoregressive language model for context-conditioned generation. *arXiv preprint arXiv:2004.07159*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Zhendong Dong and Qiang Dong. 2003. Hownet-a hybrid language and knowledge resource. In *International conference on natural language processing and knowledge engineering, 2003. Proceedings. 2003*, pages 820–824. IEEE.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional generators of words definitions. *arXiv preprint arXiv:1806.10090*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

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*, pages 2499–2509.

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, Volume 1 (Long and Short Papers)*, pages 3467–3476.

Mandar Joshi, Eunsol Choi, Omer Levy, Daniel S Weld, and Luke Zettlemoyer. 2018. pair2vec: Compositional word-pair embeddings for cross-sentence inference. *arXiv preprint arXiv:1810.08854*.

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

Cunliang Kong, Yun Chen, Hengyuan Zhang, Liner Yang, and Erhong Yang. 2022. Multitasking framework for unsupervised simple definition generation. *arXiv preprint arXiv:2203.12926*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Bin Li, Yixuan Weng, Fei Xia, Shizhu He, Bin Sun, and Shutao Li. 2022. LingJing at SemEval-2022 task 1: Multi-task self-supervised pre-training for multilingual reverse dictionary. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 29–35, Seattle, United States. Association for Computational Linguistics.

Jiahuan Li, Yu Bao, Shujian Huang, Xinyu Dai, and Jiajun Chen. 2020a. Explicit semantic decomposition for definition generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 708–717.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Ke Ni and William Yang Wang. 2017. Learning to explain non-standard English words and phrases. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 413–417, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Thirty-First AAAI Conference on Artificial Intelligence*.

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*, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

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

Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2020. Vcdm: Leveraging variational bi-encoding and deep contextualized word representations for improved definition modeling. *arXiv preprint arXiv:2010.03124*.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Aditya Srivastava and Harsha Vardhan Vemulapati. 2022. TLDR at SemEval-2022 task 1: Using transformers to learn dictionaries and representations. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 60–67, Seattle, United States. Association for Computational Linguistics.

Koki Washio, Satoshi Sekine, and Tsuneaki Kato. 2019. Bridging the defined and the defining: Exploiting implicit lexical semantic relations in definition modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3521–3527.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Liner Yang, Cunliang Kong, Yun Chen, Yang Liu, Qinan Fan, and Erhong Yang. 2019. Incorporating sememes into chinese definition modeling. *arXiv preprint arXiv:1905.06512*.

A Detailed Training Settings

Stage	Dataset	Max-epoch	Early-stop	Pooling method	λ
1	WordNet	140	40	None	0.0
	Oxford	50	10	None	0.0
	Urban	30	5	None	0.0
2	WordNet	70	40	Max	0.8
	Oxford	50	10	Max	0.8
	Urban	15	5	Max	0.8

Table 9: Detailed settings on each of our training stages, including max-epoch, early-stop threshold, pooling method and loss weight λ.

B Additional Case Study

Word	underestimate
Context	I wish people wouldn't underestimate me, or my strength, or my weakness.
Reference	regard (someone) as less capable than they really are
T5-Base	make too low an estimate of
Ours	make (someone or something) appear less important than they really are
Word	line
Context	they gave me a direct line, which was a relief, instead of those infuriating 0800 numbers
Reference	a telephone connection or service
T5-Base	a direct route
Ours	a connection item of telephone service
Word	caution
Context	a man of caution
Reference	the trait of being cautious
T5-Base	the trait of being careful
Ours	the trait of being attentive to possible danger
Word	configuration
Context	the outcome depends on the configuration of influences at the time
Reference	an arrangement of parts or elements
T5-Base	the way in which something is arranged
Ours	the arrangement of things or events in a system
Word	exercise
Context	the doctor recommended regular exercise
Reference	the activity of exerting your muscles in various ways to keep fit
T5-Base	the act of working out
Ours	the act of participating in regular physical activities

Table 10: Additional generated cases that showing the effectiveness of our method in solving the under-specific problem in definition generation.

C Perfectly Reproduced Examples

Word	net
Context	the net result
Ours	conclusive in a process or progression
Word	mysterious
Context	the new insurance policy is written without cryptic or mysterious terms
Ours	of an obscure nature
Word	legally
Context	he acted legally
Ours	in a legal manner
Word	state
Context	state your opinion
Ours	to express in words
Word	practically
Context	practically orientated institutions such as business schools
Ours	in a practical manner
Word	passionately
Context	she kissed him passionately
Ours	with passion
Word	nonprofessional
Context	the nonprofessional wives of his male colleagues
Ours	not professional
Word	buzz
Context	if you need help debugging it, you're more than welcome to give me a buzz tomorrow.
Ours	a telephone call
Word	hereafter
Context	do jews believe in the hereafter such as life after death?
Ours	life after death
Word	bop
Context	over 1,000 people bopped, jigged, jived and pogoed to some excellent bands.
Ours	dance to pop music
Word	boo bear
Context	I will love my boo bear ramero forever and always 3
Ours	pet name
Word	bang bang
Context	hey chris, do you want to bang bang tonight or will you get marcia'd?
Ours	the process of playing shoot em' up videos games with friends

Table 11: Generated cases by our method that perfectly reproduce the target definitions. Note that we omit the ground-truth reference since they are exactly the same as the generated definitions.