Definition Modelling for Appropriate Specificity

Han Huang[†], **Tomoyuki Kajiwara**[‡], **Yuki Arase**[†]
[†]Graduate School of Information Science and Technology, Osaka University
[‡]Graduate School of Science and Engineering, Ehime University
[†]{huang.han, arase}@ist.osaka-u.ac.jp
[‡]kajiwara@cs.ehime-u.ac.jp

Abstract

Definition generation techniques aim to generate a definition of a target word or phrase given a context. In previous studies, researchers have faced various issues such as the out-of-vocabulary problem and over/under-specificity problems. Over-specific definitions present narrow word meanings, whereas under-specific definitions present general and context-insensitive meanings. Herein, we propose a method for definition generation with appropriate specificity. The proposed method addresses the aforementioned problems by leveraging a pre-trained encoder-decoder model, namely Text-to-Text Transfer Transformer, and introducing a re-ranking mechanism to model specificity in definitions.[1] Experimental results on standard evaluation datasets indicate that our method significantly outperforms the previous state-of-the-art method. Moreover, manual evaluation confirms that our method effectively addresses the over/under-specificity problems.

Target	Hammer
Local Context	Health professionals are mobilising to condemn the government , propose major structural reforms , and **hammer** the ineffectual minister .
Reference	attack or criticize forcefully and relentlessly
(Ishiwatari et al., 2019)	a person who hits something
Proposed method	attack or criticize severely
Target	Bang
Local Context	Young andrew wilson , until a **bang** on the head necessitated his withdrawal , again played very well .
Reference	a sudden painful blow
(Ishiwatari et al., 2019)	(of a person) strike or strike (something) with a sudden sharp noise
Proposed method	a sudden sharp blow

Table 1: Examples of generated definitions by a previous study and our method; the previous study struggles with under- and over-specific generations.

1 Introduction

The usage of a word or phrase changes over time and new words and phrases emerge every day; therefore, the maintenance of their meanings in dictionaries is crucial but labour-intensive and time-consuming. Such definitions are also useful for computer-aided language learning (CALL), which helps language learners learn a target word or phrase (Shardlow, 2014; Srikanth and Li, 2021).

A definition generation technique aims to automatically generate a textual definition for a target word or phrase (referred to as 'target' herein) in a given sentence containing the target (referred to as 'local context' herein). Noraset et al. (2017) employed a static word embedding that models the usage of a target word or phrase, and Ni and Wang (2017), Gadetsky et al. (2018) and Ishiwatari et al. (2019) used an encoder-decoder model to generate a definition for a given sentence containing the target. However, these previous studies are limited by two problems: out-of-vocabulary (OOV) and over- and under-specificity (Noraset et al., 2017; Mickus et al., 2019; Li et al., 2020), as shown in Table 1. An under-specific definition denotes a general definition wherein part of the meaning of the target word in context is lost. In Table 1, the target word `hammer`[2] means `attack or criticize forcefully and relentlessly`, but the definition generated in the previous study failed to capture the meaning of attacking or criticizing. An over-specific definition represents a definition that contains too many details, which narrow down

[1] Code is available at https://github.com/amanotaiga/Definition_Modeling_Project

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2499–2509
November 7–11, 2021. ©2021 Association for Computational Linguistics

the meaning more than what the target truly represents. In Table 1, the target word `bang` means `a sudden painful blow`; however, the definition generated in the previous study restricts the meaning to represent `(of a person) strike or strike (something) with a sudden sharp noise`.

This study aims to automatically generate fluent definitions with appropriate specificity for a target word and phrase in a certain context. To address the aforementioned problems, we propose a re-ranking mechanism on a pre-trained encoder-decoder model. Specifically, we employ the Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020), which is a transformer-based encoder-decoder model (Vaswani et al., 2017). The pre-training with a gigantic corpus over 750GB in size effectively resolves the OOV problem. Furthermore, our re-ranking mechanism re-ranks the definitions generated from T5 based on the specificity and generality of the outputs. As presented in Table 1, our method effectively identifies a definition with appropriate specificity.

We evaluate our method on four commonly used datasets for definition generation. The results indicated that our method increased BLEU points from 2.28 to 7.95 and NIST points from 7.31 to 35.95 in comparison with those of previous state-of-the-art methods. Furthermore, a manual evaluation confirmed that our method reduced 4.0% and 0.5% of under- and over-specific definitions produced by the T5 model, respectively.

2 Related Work

An early study on definition generation (Noraset et al., 2017) proposed a method that uses pre-trained word embeddings as global contexts of a target. Owing to the lack of local contexts, this previous method cannot generate an appropriate definition for polysemous words. In contrast, Ni and Wang (2017) proposed a method that considers only the local context of a target using a word-level encoder for encoding the context to generate definitions of internet slang. The following studies considered an approach that combines the global and local contexts of a target. Gadetsky et al. (2018) proposed the first model that utilises both global and local contexts to disambiguate polysemous words. Ishiwatari et al. (2019) advanced this approach and proposed a method that models local and global contexts with multiple encoders and gate mechanisms. Washio et al. (2019) exploited lexical semantic relations between the target and words in definitions. Following Ishiwatari et al. (2019), Li et al. (2020) further introduced a module to decompose the meanings of words as discrete latent variables. Furthermore, Yang et al. (2020) established a transformer-based model for generating Chinese definitions, followed by Mickus et al. (2019), who use the attention-based model with GloVe vectors (Pennington et al., 2014) in English definition modelling.

Nevertheless, all these studies struggle with the OOV problem. Moreover, the encoder-decoder models used in these studies were trained on relatively small corpora for definition modelling. Therefore, these previous studies often result in OOV definitions, *i.e.*, ‘a target is ⟨unk⟩,’ particularly for non-standard languages (*e.g.*, internet slang). Bevilacqua et al. (2020) employed the pre-trained BART (Lewis et al., 2020) for definition generation to address the problem. Furthermore, these studies do not have any mechanism to consider the specificity of the generated definitions. Although these models succeeded in generating definitions without OOV, the generated definitions are often too general or too specific.

To this end, we employ the T5 model pre-trained on a large-scale corpora, which effectively address the OOV problem. Furthermore, we address the over/under-specific definition problem using the re-ranking mechanism.

3 Proposed Model

The overview of the proposed method is shown in Figure 1. First, we generate n-best definitions using beam search on a fine-tuned T5 model, as described in Section 3.1. Then, we obtain re-ranking scores for these definitions using two additional T5 models, as presented in Section 3.2. Specifically, we re-rank definitions based on the generation likelihood, generality, and specificity. Lastly, we assemble these scores to establish a re-ranking mechanism for identifying a definition with appropriate specificity, as described in Section 3.3.

3.1 Definition Generation with Fine-Tuned T5 Model

T5 is a unified transformer-based encoder-decoder model that is pre-trained to fill in dropped-out spans

[2] In this paper, we use `typewriter font` to present phrases in examples.

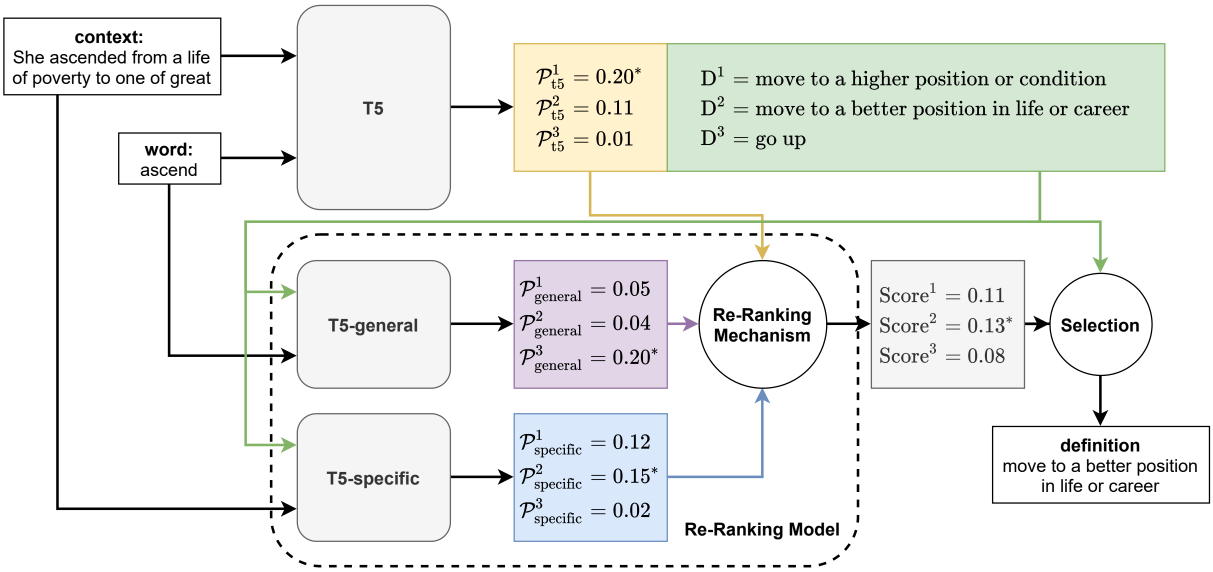
context:
She ascended from a life of poverty to one of great
word:
ascend
T5
$\mathcal{P}^1_{\text{t5}} = 0.20^*$
$\mathcal{P}^2_{\text{t5}} = 0.11$
$\mathcal{P}^3_{\text{t5}} = 0.01$
D^1 = move to a higher position or condition
D^2 = move to a better position in life or career
D^3 = go up
T5-general
$\mathcal{P}^1_{\text{general}} = 0.05$
$\mathcal{P}^2_{\text{general}} = 0.04$
$\mathcal{P}^3_{\text{general}} = 0.20^*$
Re-Ranking Mechanism
$\text{Score}^1 = 0.11$
$\text{Score}^2 = 0.13^*$
$\text{Score}^3 = 0.08$
Selection
definition
move to a better position in life or career
T5-specific
$\mathcal{P}^1_{\text{specific}} = 0.12$
$\mathcal{P}^2_{\text{specific}} = 0.15^*$
$\mathcal{P}^3_{\text{specific}} = 0.02$
Re-Ranking Model

Figure 1: The proposed method consists of a definition generator and two re-rankers for controlling specificity.

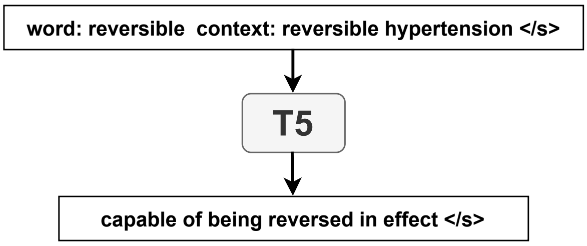
word: reversible context: reversible hypertension </s>
T5
capable of being reversed in effect </s>

Figure 2: Definition generation by T5 model, which takes a target word or phrase and a local context as input, and outputs n-best definitions.

of text. It is trained on a large-scale corpus scraped from the web combined with corpora for supervised tasks of translation, summarisation, classification, and reading comprehension. T5 can handle various text-based language problems in natural language processing after fine-tuning.

We follow the fine-tuning procedure described in Raffel et al. (2020), as shown in Figure 2. First, we prepare the pairs of targets and the corresponding local contexts. Second, we concatenate them with the labels, 'word:' and 'context:'. Then, we input them into the encoder of T5 after sub-word segmentation by SentencePiece (Kudo and Richardson, 2018) and train the model to generate definitions using the cross-entropy loss. Through this fine-tuning, T5 learns to generate the definition of the target conditioned in the local context.

Generation Likelihood For re-ranking, we consider the generation likelihood of each definition. Given a target w^* and corresponding local context C, the fine-tuned T5 model predicts the probability of words in the output $D = \{w_1, \ldots, w_T\}$, which can be formulated using a conditional language model:

$$P(D \,|\, C, w^*) = \prod_{t=1}^{T} P(D_t \,|\, D_{<t}, w^*, C). \quad (1)$$

For each output, we obtain the generation likelihood $\mathcal{P}_{\text{T5}}$ for re-ranking:

$$\mathcal{P}_{\text{T5}} = -\log(P(D \,|\, C, w^*)). \quad (2)$$

The lower the score, the corresponding definition is more likely to be generated.

3.2 Re-Ranking Models

To identify a definition with appropriate specificity, we use two estimators: one evaluates the level of over-specificity of a definition and the other evaluates the level of under-specificity. In the quality estimation of machine translation, force-decoding has been used to estimate the likelihoods of machine translation outputs, achieving state-of-the-art performance (Thompson and Post, 2020). Inspired by this approach, we fine-tune other T5 models and use force-decoding for estimating the levels of over/under-specificity.

Over-Specificity We observed that over-specific definitions are generated when a generation model is overly affected by local contexts, *i.e.*, the generated definitions tend to contain words that are relevant to those in the local context. For example,

	WordNet			Oxford			Urban			Wikipedia		
	Train	Valid	Test	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	151,995	8,361	8,397
Entries	13,883	1,752	1,775	97,855	12,232	12,232	411,384	57,883	36,450	887,455	44,003	57,232
Context length	5.81	5.64	5.77	17.74	17.80	17.56	10.89	10.86	11.22	18.79	19.21	19.02
Desc. length	6.61	6.61	6.85	11.02	10.99	10.95	10.99	10.95	12.05	5.89	6.31	6.94

Table 2: Statistics of dataset

the specific definition of `bang` in Table 1 contains the phrases, `of a person` and `strike` that are likely to be affected by the phrases, `andrew wilson` (person name) and `play` in the local context, respectively. Based on this observation, we assume that an over-specific definition results in a higher probability of force-decoding the local context.

We first fine-tune a T5 model to generate a local context conditioned on a definition (reference) and use it as specificity estimator. We force-decode the local context C conditioned on a generated definition D. The specificity score $\mathcal{P}_{\text{specific}}$ can be represented as follows:

$$\mathcal{P}_{\text{specific}} = -\log(P(C \mid D)). \tag{3}$$

The lower the score, the more specific the generated definition.

Under-Specificity In contrast to over-specific definitions, we observed that excessively general definitions are overly affected by the most common meaning of a target word and ignore the local contexts. For example, the excessively general definition of `Hammer` in Table 1 represents the most common meaning of the target without considering the local context. Based on this observation, we assume that an under-specific definition can be easily force-decoded from the target.

We fine-tune another T5 model to generate a definition conditioned on a target without a local context as the generality estimator. Given a target w^*, the generality estimator force-decodes the definition D. The under-specificity score $\mathcal{P}_{\text{general}}$ can be represented as follows:

$$\mathcal{P}_{\text{general}} = -\log(P(D \mid w^*)). \tag{4}$$

The lower the score, the more general the generated definition.

3.3 Combining Re-Ranking Scores

Finally, we combine the generation likelihood $\mathcal{P}_{\text{T5}}$, over-specificity score $\mathcal{P}_{\text{specific}}$, and under-specificity score $\mathcal{P}_{\text{general}}$ to re-rank n-best definitions generated by T5.[3] We use a simple linear combination of these scores as:

$$r = \alpha\mathcal{P}_{\text{specific}} + \beta\mathcal{P}_{\text{general}} + (1 - \alpha - \beta)\mathcal{P}_{\text{T5}}, \tag{5}$$

where α and β are hyper-parameters ranging from 0 to 1. The values of α and β are tuned using development sets. The n-best definitions are re-ranked based on the values of r, and top-1 is output as a definition.

4 Experimental Setup

We compared the performance of the proposed method with those of previous state-of-the-art methods using the standard datasets for automatic definition generation. This section describes the experimental setup in detail.

4.1 Evaluation Datasets

We used four evaluation datasets created in previous studies (Ni and Wang, 2017; Noraset et al., 2017; Gadetsky et al., 2018; Ishiwatari et al., 2019), which were assembled by Ishiwatari et al. (2019).[4]

Table 2 shows the statistics of these datasets. Each entry in a dataset consists of three elements: (1) a target word or phrase, (2) a corresponding definition of the target, and (3) one usage example of the target as a local context. It is noteworthy that if a target has multiple definitions and local contexts, we treat them as different entries (Ishiwatari et al., 2019).

Wordnet dataset The Wordnet dataset was collected from entries of the GNU Collaborative International Dictionary of English[5] and Wordnet's glosses (Miller, 1995) by Noraset et al. (2017). The original dataset provides only a target and its definition. This dataset was expanded by Ishiwatari et al.

[3] Note that we select the definition with the minimum score (negative log-likelihoods)

[4] `http://www.tkl.iis.u-tokyo.ac.jp/~ishiwatari/naacl_data.zip`

[5] `http://wwwgcide.gnu.org.ua`

	WordNet			Oxford		
	BLEU	NIST	OOV rate	BLEU	NIST	OOV rate
(Gadetsky et al., 2018)	23.77	44.30	0%	17.45	35.79	13.47%
(Ni and Wang, 2017)	24.78	40.32	0%	17.58	31.30	17.92%
(Noraset et al., 2017)	23.59	49.70	0%	14.95	32.79	17.02%
(Ishiwatari et al., 2019)	25.19	43.54	0%	18.57	38.22	15.66%
Proposed method	**32.72***	**64.57***	0%	**26.52***	**74.17***	0%
T5-base	31.72	57.35	0%	25.44	66.92	0%
T5+specific score	30.61	59.16*	0%	26.10*	73.03*	0%
T5+general score	32.32	64.08*	0%	26.00*	68.50*	0%

	Urban			Wikipedia		
	BLEU	NIST	OOV rate	BLEU	NIST	OOV rate
(Gadetsky et al., 2018)	8.81	19.43	77.00%	44.96	33.17	3.40%
(Ni and Wang, 2017)	8.99	17.39	80.88%	52.69	55.25	2.40%
(Noraset et al., 2017)	5.15	10.45	90.22%	44.59	33.45	7.24%
(Ishiwatari et al., 2019)	9.93	19.29	78.90%	53.38	56.69	3.82%
Proposed method	**17.71***	**35.53***	0%	**55.61**	**64.00***	0%
T5-base	17.66	26.86	0%	**55.66**	60.86	0%
T5+specific score	17.15	32.48*	0%	55.40	63.65*	0%
T5+general score	17.51	32.04*	0%	55.39	60.76	0%

Table 3: Scores of BLEU and NIST on test sets (*: Scores are significant at $p < 0.05$ between proposed method and T5-base.)

(2019) to add usage examples for each entry and remove the entries that have no usage examples. We used the expanded version of the dataset for fair comparison to previous studies that use local contexts for definition generation.

Oxford dataset The Oxford dataset was collected using APIs of Oxford Dictionaries[6](2018) by Gadetsky et al. (2018).

Urban dataset The Urban dataset is a collection from the non-standard English corpus from Urban Dictionary (UD)[7], which is the largest online slang dictionary collected by Ni and Wang (2017). In this dataset, all terms, definitions, and examples are submitted by internet users. Unlike the Wordnet and Oxford datasets, the Urban dataset contains not only words but also phrases. We noticed that this dataset contains erroneous entries whose definitions are single Arabic numerals or part-of-speech tags. We excluded these erroneous entries from evaluation using a simple heuristic.

Wikipedia dataset The Wikipedia dataset was collected from Wikipedia[8] and Wikidata[9] by Ishiwatari et al. (2019). The Wikipedia dataset also provides phrases as targets, but their domains are across different fields, whereas phrases in the Urban dataset are all online slangs.

4.2 Evaluation Metrics

Following the previous studies, we used BLEU (Papineni et al., 2002) as an automatic evaluation metric. However, BLEU is vulnerable to the evaluation of definition generation because the references are short (less than 12 words as shown in Table 2) and many of them have prototypical expressions, such as '*the quality of being* something'. Moreover, we found that definitions generated by previous studies have high OOV rates, which is critical in definition generation. Although definitions of high OOV rates, such as 'the quality of being ⟨unk⟩', are inefficient, BLEU evaluates them highly because

[6] https://developer.oxforddictionaries.com/

[7] https://www.urbandictionary.com/

[8] https://dumps.wikimedia.org/enwiki/20170720/

[9] https://dumps.wikimedia.org/wikidatawiki/entities/20170802/

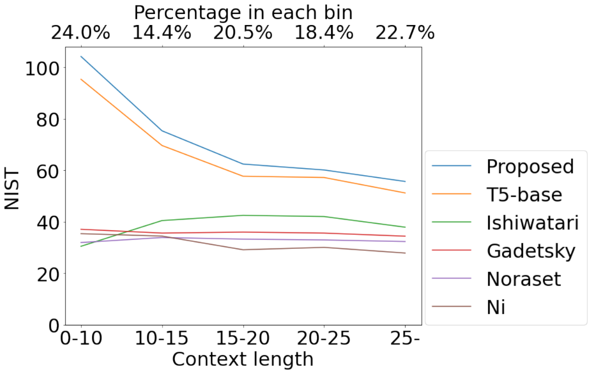

Figure 3: Effect of length of local contexts

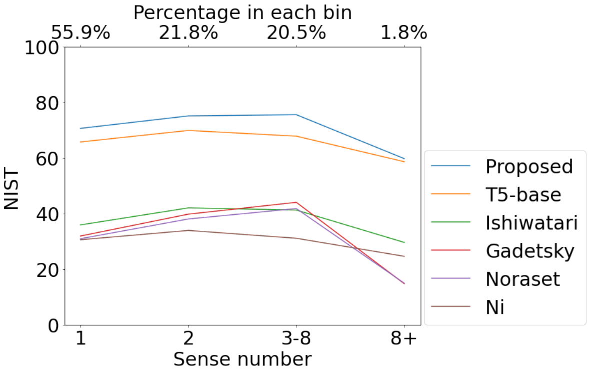

Figure 4: Effect of number of senses in target words

the majority of words matches with those of the reference.

To address this issue, we used NIST (Doddington, 2002) and OOV rate as evaluation metrics to properly evaluate the quality of generated definitions. NIST focuses on content words by giving more weightage to them. This makes NIST more informative than solely assigning an equal weight to each m-gram as in BLEU.

Herein, we report on the results of statistical significance testing. We apply the Wilcoxon signed-rank test (Wilcoxon, 1945), which tests the null hypothesis that two related paired samples are from the same distribution.

4.3 Implementation Details

We compared our method to a previous state-of-the-art method (Ishiwatari et al., 2019), as well as representative methods of definition generation (Ni and Wang, 2017; Noraset et al., 2017; Gadetsky et al., 2018). We replicated experiments using implementations released by Ishiwatari et al. (2019).[10] While these previous studies use word2vec[11] as global contexts, the vocabulary coverage of Wordnet, Oxford, Urban, and Wikipedia is 100%, 83%, 21%, and 27%, respectively, as reported by Ishiwatari et al. (2019). As an ablation study, we also compared our method to a simply fine-tuned T5 model without the re-ranking mechanism, as well as re-ranking models with only over/under-specificity scores.

For implementing the proposed method and its variants, we used T5-base[12] that has 220 million parameters in 12-layers of transformer blocks, consisting of 768 hidden-states, 3,072 feed-forward hidden-states, and 12-heads for multi-head attention. We fine-tuned T5-base on each evaluation dataset using Adam (Kingma and Ba, 2015) as an optimiser with a constant learning rate of 0.0003 and a batch size of 16. Fine-tuning was terminated when the value of cross-entropy loss measured on a validation set stopped decreasing for 5 continuous epochs.

During evaluation, the model generated 100 outputs for each input through beam search. To determine the best weights of α and β in Equation (5), we performed a grid search on each validation set. We set these hyper-parameters to maximise BLEU and NIST, respectively. Note that we reported the BLEU and NIST scores measured on test sets as evaluation results, where hyper-parameters were tuned for each metric. Although the best values are dataset dependent, setting α in the range of $[0.4, 0.8]$ and β in the range of $[0.1, 0.4]$ consistently performed well. The values of α and β for all cases are listed in the appendices.

5 Experimental Results and Analyses

We present the results of the automatic evaluation and further conduct quantitative analyses to declare the performance of the proposed method under different conditions.[13] Furthermore, we conduct a manual analysis to investigate whether the over/under-specific definition problems are addressed.

[10] https://github.com/shonosuke/ishiwatari-naacl2019

[11] https://code.google.com/archive/p/word2vec/

[12] https://huggingface.co/t5-base

[13] We further present experimental results for relevant comparisons with Washio et al. (2019), Li et al. (2020), and Bevilacqua et al. (2020) in the appendices.

Error type	Word	Definition
(1) Over-specified	waft	(of an unpleasant smell) spread through the air
(2) Self-reference	self-consciousness	the state of being self-conscious
(3) Wrong part-of-speech	red-hot	of the most recent interest or importance
(4) Under-specified	forerunner	a thing that precedes another
(5) Opposite	hollow	a cavity that is felt by food
(6) Similar semantics	machine	a device with automatic functions
(7) Incorrect	first	the next after all others in a set of items
(8) Correct	winery	a factory or business that produces wine

Table 4: Definition of error types

	T5-base	Proposed method
(1) Over-specified	5.5%	5.0%
(2) Self-reference	3.0%	3.5%
(3) Wrong part-of-speech	1.0%	1.0%
(4) Under-specified	9.0%	5.0%
(5) Opposite	1.0%	1.0%
(6) Similar semantics	37.0%	34.5%
(7) Incorrect	25.5%	20.0%
(8) Correct	36.5%	45.0%

Table 5: Percentage of errors in generated definitions by the fine-tuned T5 and our method

5.1 Experimental Results

Table 3 presents the BLEU and NIST scores for all the compared models measured on the test sets.[14] The results indicate that the proposed method consistently outperforms the four baselines in all datasets by a large margin on BLEU and NIST.

The higher performance of the proposed method on NIST indicates that it can generate proper content words compared to those in the baseline methods. Moreover, the performance gaps between our method and the strongest baseline methods on Oxford and Urban Dictionary are larger (35.95 and 16.1 points, respectively) than those on Wikipedia and Wordnet (7.31 and 14.87 points, respectively), although the former datasets are more challenging due to the longer average length of the definitions.

A large portion of the words and phrases in the Urban Dictionary dataset is not available in word2vec, thereby restricting the global contexts in the baseline models. Our method achieves a high NIST score even for the Urban Dictionary dataset, which has been considered excessively difficult in the state-of-the-art method (Ishiwatari et al., 2019). This result indicates that the proposed method is robust against the OOV problem.

5.2 Ablation Study

Our re-ranking method outperforms the strong T5-base model on Wordnet, Oxford, and Urban datasets on BLEU, and all datasets on NIST. T5+specific score achieves a higher NIST than that of T5-base on four datasets, which shows that the model tends to generate under-specified definitions. For the Wordnet dataset, the general score (T5+general score) is more beneficial than the specific score. This is because the average context length is the shortest among the four datasets, which implies that the specificity of the contexts is lesser than that in the other three datasets. The proposed method achieves the highest performance by combining the general and specific scores.

5.3 Quantitative Analysis

Intuitively, the length of local contexts and the number of senses of the target are the primary factors that affect the definition generation quality. With regard to the former, longer contexts are more difficult to encode to properly represent their meanings. With regard to the latter, targets with a larger number of senses are more difficult to determine the sense that is represented in the local context.

For analysing these factors, we use the Oxford dataset because it contains different types of targets with relatively longer local contexts, as shown in Table 2. Figure 3 shows the NIST scores on different lengths of local contexts. T5 and the proposed method achieve significantly higher NIST scores across different lengths of local contexts. This can be attributed to the powerful encoder pre-

[14] Note that these BLEU scores are slightly different from those reported by Ishiwatari et al. (2019), likely owing to the differences in computational environments.

Target	Ascend	Electronic
Context	She **ascended** from a life of poverty to one of great	1987 was ... for **electronic** dance music .
Reference	move to a better position in life or to a better job	(of music) produced by electronic instruments
Gadetsky	move or move upward	to or denoting the ⟨unk⟩ of a ⟨unk⟩
Ishiwatari	go up	relating to or denoting the branch of science concerned with the ⟨unk⟩ of ⟨unk⟩ and ⟨unk⟩
T5-base	move to a higher position or condition	relating to or using electronics
Ours	move to a better position in life or career	denoting or relating to music produced by electro-mechanical means
Target	Debut	Cry
Context	... he began working professionally , **debuting** at the gaiety theatre ...	she **cried** bitterly when she heard the news ... .
Reference	perform in public for the first time	shed tears because of sadness , rage , or pain
Gadetsky	a person who is ⟨unk⟩ or ⟨unk⟩	a loud utterance
Ishiwatari	a person 's first ⟨unk⟩	make a loud , loud sound
T5-base	make one's first appearance	utter emotions such as sorrow or pain
Ours	perform for the first time in public	shed tears because of a strong emotion
Target	Acquire	Worker
Context	Children **acquire** language at an amazing rate	The guy is a **worker** , there 's no doubt he 's a worker .
Reference	gain knowledge or skills	a person who works hard
Gadetsky	take (something) into a particular place	a person who is employed to do something
Ishiwatari	be ⟨unk⟩	a person who works in a specified way
T5-base	the ability to recognize or learn a language	a person who does manual or other work for wages
Ours	the ability to learn knowledge or skills	a person who works hard

Table 6: Examples of generated definitions for words sampled from the Oxford dataset and Wordnet

trained on a large-scale text corpus. The proposed method even outperforms T5 owing to the effective re-ranking mechanism.

Figure 4 shows the impact of the number of senses of targets. It is reasonable that the method proposed by Noraset et al. (2017) performs poorly because it considers only global contexts, *i.e.*, word embeddings. Our method consistently outperforms all these previous methods on any numbers of target senses.

5.4 Error Analysis

As there is no means to automatically evaluate methods for the over- and under-specificity problems, we conducted a manual error analysis. We randomly sampled 200 generated definitions by T5-base and the proposed method from the Oxford dataset. For the error type, we followed Noraset et al. (2017), where we added the 'over-specified' definition.[15] We provide an example for each error type in Table 4.

Table 5 shows the distribution of errors in definitions generated by T5-base and the proposed method. Overall, our method reduces the errors of T5-base for most error types, resulting in the generation of 8.5% more correct definitions (type (8)) than that of the strong T5-base model. The proposed method exhibits a larger improvement for the under-specificity problem (type (4)) than that of T5-base, and the error rate of the proposed method is 4% lower than that of the T5-base model. The improvement can be attributed to the estimation of the degree of under-specificity.

For the over-specificity problem (type (1)), the error rate of the proposed method is 0.5% lower than that of T5-base. This is because if the prediction generated by T5-base is over-specific, other n-nest predictions also tend to be over-specific in certain aspects. This causes our re-ranking model to have a lesser chance of selecting more general predictions.

5.5 Examples

Table 6 presents examples of generated definitions by Ishiwatari et al. (2019), Gadetsky et al. (2018), T5-base, and the proposed method sampled from the Oxford dataset. Evidently, the previous methods face the OOV problem by generating unknown words (⟨unk⟩) frequently. Furthermore, these methods generate under-specific definitions for `ascend` and `cry` an over-specific definition for `worker`.

[15] We also removed 'overusing common phrase' because this phenomenon was barely observed.

In contrast, both T5-base and the proposed method generate fluent definitions for all targets. For the target `ascend`, the meaning in the local context represents `away from a bad situation in life`. The definition generated by T5-base is too general, where `better position in life` is more appropriate than `higher position`. Similarly, T5-base generates under-specific definitions for `debut`, `electronic` and `cry`, whereas the proposed method generates appropriate definitions.

For the target word `acquire`, although the word `language` appears in the local context, it is too narrow to define this word in association with language learning, as in the T5-base output. Similarly, the T5-base definition of `worker` is also over-specific. Only the proposed method generates definitions with appropriate specificity for these targets.

6 Conclusion

We addressed the definition generation problem and developed a re-ranking mechanism equipped with a pre-trained T5 model. The quantitative and qualitative analyses confirmed that the proposed method significantly outperformed previous state-of-the-art methods and the strong fine-tuned T5 model and successfully generated definitions with appropriate specificity. As future work, we aim to investigate the effectiveness of the proposed method for cross-lingual definition generation.

Acknowledgement

The authors sincerely appreciate Shonosuke Ishiwatari for clarifying the implementation and evaluation details of their work. We also thank the anonymous reviewers for their valuable comments. This work was supported by JSPS KAKENHI Grant Number 21H03564.

References

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 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, Online. Association for Computational Linguistics.

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*, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 266–271, Melbourne, Australia. Association for Computational Linguistics.

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 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3467–3476, Minneapolis, Minnesota. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

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 Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Timothee Mickus, Denis Paperno, and Matthieu Constant. 2019. Mark my word: A sequence-to-sequence approach to definition modeling. In *Proceedings of the NLPL Workshop on Deep Learning for Natural Language Processing*, pages 1–11, Turku, Finland. Linköping University Electronic Press.

George A. Miller. 1995. WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Ke Ni and William Yang Wang. 2017. Learning to explain non-standard English words and phrases. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 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 *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 3259–3266.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Natural Language Processing 2014*, 4(1).

Neha Srikanth and Junyi Jessy Li. 2021. Elaborative simplification: Content addition and explanation generation in text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, pages 6000–6010.

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 Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3521–3527, Hong Kong, China. Association for Computational Linguistics.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

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*, 28:1669–1677.

A Values of Re-ranking Parameters

	Proposed method	
	BLEU	NIST
Wordnet	$[0.4, 0.4, 0.2]$	$[0.6, 0.2, 0.2]$
Oxford	$[0.4, 0.2, 0.4]$	$[0.8, 0.1, 0.1]$
Urban	$[0.3, 0.7, 0.0]$	$[0.7, 0.2, 0.1]$
Wikipedia	$[0.5, 0.1, 0.4]$	$[0.6, 0.1, 0.3]$

Table 7: Hyper-parameters tuned for BLEU and NIST metrics on validation sets, presented in the format of $[\alpha, \beta, 1-\alpha-\beta]$

To derive the best parameters on BLEU and NIST metrics for each dataset, we applied grid search on the validation set. The range of the grid search was $[0, 1]$ with a step of 0.1. The weights on each dataset are presented in Table 7. In the table, the weights follow the format $[\alpha, \beta, 1-\alpha-\beta]$, where α weighs the over-specificity score $\mathcal{P}_{\text{specific}}$, β weighs the under-specificity score $\mathcal{P}_{\text{general}}$, and the last weighs the generation likelihood $\mathcal{P}_{\text{T5}}$.

B Additional Experimental Results

We present the experimental results in this section for some relevant comparisons that not reported in the main text. For the same dataset, our BLEU score varies for different calculation methods. All the BLEU scores of previous studies are borrowed from the original papers.

The comparison of the obtained result with Li et al. (2020) is shown in Table 8. They used the

	WordNet BLEU	Oxford BLEU
Li et al. (2020)	26.48	20.86
Proposed method	**32.72**	**26.52**

Table 8: Scores of sentence BLEU with multi-references on Wordnet and Oxford dataset

	Oxford BLEU
Washio et al. (2019)	12.3
Proposed method	**23.26**

Table 9: Scores of sentence BLEU with single-reference on Oxford dataset

	Oxford BLEU
Bevilacqua et al. (2020)	9.9
Proposed method	**10.15**

Table 10: Scores of corpus BLEU on Oxford dataset

average of sentence BLEU with multi-references based on Ishiwatari et al. (2019). The comparison of the obtained result with Washio et al. (2019) is shown in Table 9. They used the average of sentence BLEU with single-reference based on Gadetsky et al. (2018). The comparison of the obtained result with Bevilacqua et al. (2020) is shown in Table 10. They used the corpus BLEU calculated by sacreBLEU script (Post, 2018). The proposed method outperforms all of these previous studies.