AUTOGRAPHEX: Zero-shot Biomedical Definition Generation with Automatic Prompting

Anonymous ACL submission

Abstract

Describing terminologies with definition texts is an important step towards understanding the scientific literature, especially for domains with limited labeled terminologies. Previous works have sought to design supervised neural text generation models to solve the biomedical terminology generation task, but most of them failed to define never-before-seen terminologies in newly emerging research fields. Here, we tackle this challenge by introducing a zero-shot definition generation model based on *prompting*, a recent approach for eliciting knowledge from pre-trained language models, with automatically generated prompts. Furthermore, we enhanced the biomedical terminology dataset by adding descriptive texts to each biomedical subdiscipline, thus enabling zero-shot learning scenarios. Our model outperformed existing supervised baseline and the baseline pre-trained language model that employs manually crafted prompts by up to 52 and 6 BLEU score, respectively.

1 Introduction

Describing new terminologies has become a task of great significance in scientific research, as expert curated definitions cannot scale to the rapidly emerging terminologies, especially in new research topics(Cimino et al., 1994). Prior works have sought to generate biomedical definitions via neural text generation models(Liu et al., 2021b), leveraging both the terminology text and the terminology relation graph. However, these methods focus on designing supervised learning models by assuming the availability of sufficient annotated terminology text in every scientific subdiscipline, which is seldom the case. In many newly emerging research topics, such as COVID-19, people have very few expert curated definitions, hindering the usage of those fully supervised learning models(Baines and Elliott, 2020). On the other hand, large amounts of gold definitions are available in some other research domains, and descriptive texts of scientific subdisciplines are widely accessible.

To form a more realistic setting, we propose our task as zero-shot definition text generation, similar to Zero-shot text classification (ZSC) to classify text using label descriptions without any examples(Yin et al., 2019). In the past, lines of few-shot text generation models have been proposed to address this task(Lin et al., 2019; Song et al., 2020; Schick and Schütze, 2021); however, most of these models fail to fully leverage the language models pre-trained on massive amounts of raw text. Recently, *prompting* has become a popular approach among the NLP community to elicit knowledge from large language models, allowing for direct performance of few-shot and zero-shot learning(Seoh et al., 2021; Gao et al., 2021; Brown et al., 2020; Liu et al., 2021a). For instance, text summarization can be formalized as a language model task by adding "TL; DR" to the end of an article(Radford et al., 2019). Unfortunately, it is challenging to manually acquire prompts, and these prompts are likely to be sub-optimal. Hence, people have introduced automatic prompts generating tools in order to overcome the need of human crafted prompts template(Shin et al., 2020).

In this paper, we propose a zero-shot biomedical definition generation dataset GRAPHINE-ZERO and a biomedical definition generation model AUTOGRAPHEX. An overview of our method is shown in Figure 1. We introduced GRAPHINE-ZERO by removing the intersected terminologies of independent graphs in GRAPHINE and collecting dataset descriptions from biomedical ontology databases for each graph. Our model, AUTOGRAPHEX, generates definitions in a particular biomedical subdiscipline without any training data. Specifically, AUTOGRAPHEX only leverages descriptive texts in the target subdiscipline and expert-curated definition texts in other biomedical domains. Given a pre-trained language model, it appends prompts

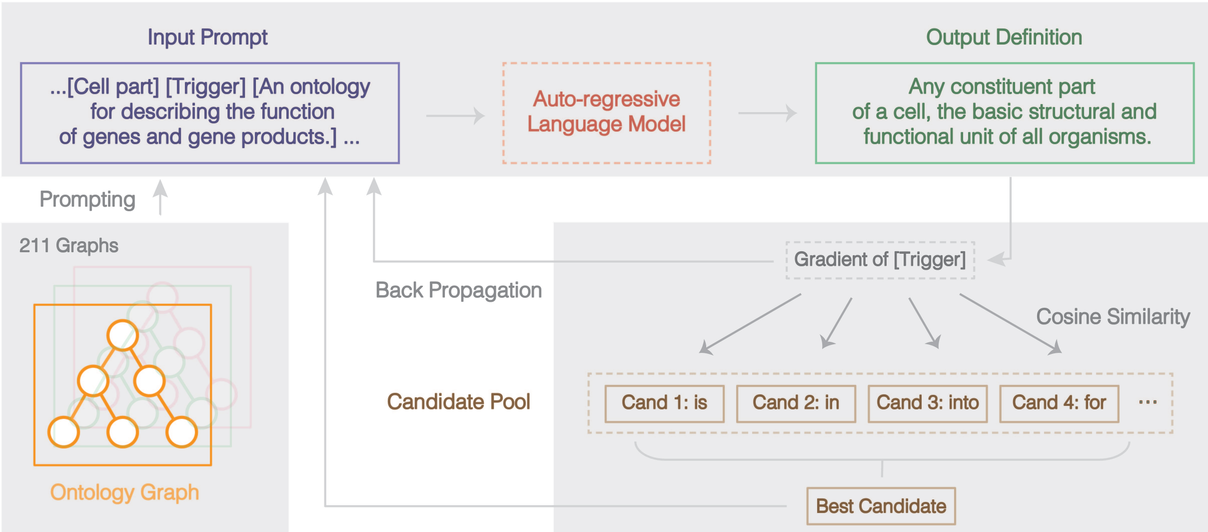

Figure 1: An overview of AUTOGRAPHEX. The input terminology text from an ontology graph is first transformed into prompts based on the template, which serve as input to the language model to generate the definition text. AUTOGRAPHEX selects candidate token based on the cosine similarity between its embedding vector and the gradient with respect to trigger token. This process is carried out iteratively to select the best candidate tokens.

that consist of the dataset description, terminology text, neighbour biomedical ontology regarding the graph and a collection of trigger tokens. Furthermore, AUTOGRAPHEX uses a gradient-based prompts search method to automatically select subwords for predicting those trigger tokens. Besides, we validate the effectiveness of AUTOGRAPHEX with several machine translation metrics including BLEU, METEOR and NIST. Concretely, AUTOGRAPHEX outperformed supervised baseline, GRAPHEX, by 52.32, 73.18 and 14.47 in terms of BLEU score, METEOR score and NIST score (Table 2).

2 Methods

2.1 Dataset: GRAPHINE-ZERO

Graphine dataset consists of 2,010,648 biomedical terminology definition pairs encapsulated in 227 directed acyclic graphs (DAGs)(Liu et al., 2021b). Each edge in the DAGs represents an is-a relation between two nodes. We created our dataset GRAPHINE-ZERO based on GRAPHINE by removing the intersected terminologies in different graphs, leaving 211 DAGs, to set different DAGs as independent tasks in zero-shot learning. We further obtain descriptions for each DAG from ontology databases, Open Biological and Biomedical Ontology Foundry (OBO)(Smith et al., 2007), BioPortal and EMBL-EBI Ontology Lookup Service (OLS)(Noy et al., 2009; Jupp et al., 2015). For example, the description for GO (Gene Ontology) dataset is: *"An ontology for describing the function of genes and gene products."*

2.2 Task: zero-shot definition generation

Let $\mathcal{G} = \{G_1, G_2, ...G_N\}$ denote N graphs. For each graph $G = \{\mathcal{V}, \mathcal{E}, \mathbf{X}, \mathbf{Y}, \mathbf{d}\}$, $\mathcal{V}$ is a set of nodes, $\mathcal{E}$ is a set of edges, $\mathbf{d}$ is the description sentence for the graph G. Each node is associated with a terminology $x_i \in \mathbf{X}$ and a definition $y_i \in \mathbf{Y}$. Both x_i and y_i are token sequences. We then split the meta-set $\mathcal{G}$ into $\mathcal{G}_{train}$, $\mathcal{G}_{valid}$ and $\mathcal{G}_{test}$. A zero-shot definition generation model is trained on $\mathcal{G}_{train}$ and $\mathcal{G}_{valid}$ to adapt to the task of definition generation in $\mathcal{G}_{test}$ with no definition samples in each graph $G \in \mathcal{G}_{test}$.

2.3 Definition generation with prompting

Definition generation problems can be solved in the framework of using prefix prompts together with pre-trained auto-regressive language models such as GPT and BART(Lewis et al., 2019). In our task, the prefix prompt should consider information of terminology text, neighbor ontology and description text. The prompting text should be similar to what one would typically write and in accordance with the real biomedical facts corresponding to ontology hierarchy. We came up with the following manually crafted prompt in Table 1, in which [term] is the terminology text, [def] is the definition text of the biomedical ontology and

Methods	$\mathbf{X_{prompt}}$	$\mathbf{Y_{prompt}}$
Manual	[Term] is in [Desc]. Concat([Term] is a [Parent-1],...[Term] is a [Parent-m]). Concat([Child-1] is a [Term],...[Child-n] is a [Term]). The definition of [Term] is	[Def]
AutoGraphex	[Term] [T] [T] [Desc]. Concat([Term] [T] [T] [Parent-1],...[Term] [T] [T] [Parent-m]). Concat([Child-1] [T] [T] [Term],...[Child-n] [T] [T] [Term]). [T] [T] [T] [Term] [T]	[Def]

Table 1: Manually crafted prompt templates and the prompt templates used by AUTOGRAPHEX.

[desc] is the description text of the DAG that the biomedical ontology belongs to. Each ontology node has several parent nodes , indicating the current node belongs to its parent ontology node, and several child nodes that belong to the ontology of the current node. We use [parent-i] and [child-i] to represent the terminology text of parent nodes and child nodes. Formally, our pre-trained auto-regressive language model optimizes the following loss $\mathcal{L}$ in the training stage:

$$\mathcal{L} = \prod_{x \in X_p, y \in Y_p} p_M(y|x) \quad (1)$$

where X_p and Y_p are the prompt dataset after applying the templates on the training dataset, and p_M is the language model loss.

The pre-trained weights of large auto-regressive models are available, but these models are not pre-trained on general biomedical corpus. People have proposed masked language models and other representation learning models pre-trained on scientific publications(Gu et al., 2020; Lee et al., 2020b; Beltagy et al., 2019; Cohan et al., 2020); however, these models are not appropriate for text generation tasks. To tackle this challenge, we trained our language model on the training split $\mathcal{G}_{train}$ and the validation split $\mathcal{G}_{valid}$ of the meta-set using the manually crafted prompt mentioned above. In the test stage, we remove the [def] sentence and let the language model generate sentences conditioned on the prompt. We compared this prompt with several other choices and achieved best performance among them. However, the possible search space for prompts is large, and we were only able to test a few comparison prompt templates. Hence, we rely on automatic prompt search to address the issue next.

2.4 Model: AUTOGRAPHEX

The idea of AUTOGRAPHEX is to first create prompt templates with trigger tokens, and then fill the trigger tokens with real subword tokens that maximize the likelihood on training dataset. Based on the automatic prompts searching work on Masked Language Models (MLMs)(Shin et al., 2020), we employ a gradient-based subword search strategy on the auto-regressive model. First, we construct the following template based on the prompt template proposed in subsection 2.3 by replacing manually crafted words with trigger tokens in Table 1. We use [T] to represent trigger tokens in the template, and initialize these tokens with the words in manually created templates. These trigger tokens are then updated iteratively by swapping trigger tokens with the tokens in the vocabulary. We select the candidate token through maximizing the cosine similarity of the candidate tokens embedding and the gradient vector with respect to the trigger token embedding. Formally, we have:

$$\mathcal{V}_{cand} = max_{\omega \in \mathcal{V}}(\omega^T \nabla_{\omega_t} \log(p(\mathbf{x}_{prompt}))) \quad (2)$$

where ω is the embedding of the candidate token, ω_t is the gradient of log-likelihood loss function with respect to the embedding vector of the trigger tokens, p is the language model loss and $\mathbf{x}_{prompt}$ is the prompt text. Then we replace the trigger token with the candidate token and calculate the language model loss to decide whether to use the token or not. After several iterations, all of the candidate tokens are fixed and this prompt is used for definition generation in the test stage.

3 Experimental Results

3.1 Baselines

We compared our methods with G-META (Huang and Zitnik, 2020) and GRAPHEX (Liu et al., 2021b).

G-META is a meta-learning algorithm for graphs. It uses local subgraphs to transfer subgraph-specific information and learn transferable knowledge faster

Model	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	NIST
G-meta	20.21	15.36	13.24	9.82	7.63	0.21
Graphex(Supervised)	34.35	26.97	22.99	20.21	16.57	1.15
Prompting	69.85	69.02	67.13	66.37	87.94	11.94
AutoGraphex	**74.63**	**74.09**	**73.59**	**72.53**	**90.25**	**15.62**

Table 2: Comparison of the zero-shot definition generation performance of AUTOGRAPHEX against baselines in terms of BLEU, METEOR and NIST score.

via meta gradients based on MAML(Finn et al., 2017). Note that G-META cannot be used on text generation task directly, so we implemented a transformer version of G-META under its framework as G-META is model-agnostic (Vaswani et al., 2017).

GRAPHEX is a supervised graph-aware definition generation approach. It first calculates the global semantic embedding through propagating terminology and definition on the graph, and then obtains a local embedding of the specific terminology. GRAPHEX uses transformer as its text generation model and uses BIOBERT as its pre-trained text encoder(Lee et al., 2020b).

3.2 Experimental Setup

We compared our methods and baselines methods on GRAPHINE-ZERO. We used 118 DAGs as the training dataset, 22 DAGs as the validation dataset and 71 DAGs as the test dataset. As we removed similar terminology in different DAGs, GRAPHINE-ZERO has only 1,366,064 terminology definition pairs. The smallest 5 DAGs only include 17, 40, 41, 45 and 92 terminology definition pairs, while the largest 5 DAGs includes 34,002, 43,795, 114,062, 121,610 and 321,860 pairs, indicating the data quantity in different biomedical subdisciplines are severely unbalanced. Meta-learning and few-shot learning methods may fail to generalize well in unbalanced and out-of-distribution data regimes(Lee et al., 2020a).

We used several machine translation metrics for performance comparison. We used six standard metrics including **BLEU1-4**, **METEOR** and **NIST** (Papineni et al., 2002) (Banerjee and Lavie, 2005) (Doddington, 2002). **BLEU** measures the n-gram similarities between generated and reference sentences. **METEOR** considers synonyms when comparing unigram and using F1 score instead of precision used in **BLEU**. **NIST** re-weights words by frequency when matching n-gram overlap to adjust the contribution of common words.

We selected BART as our pre-trained autoregressive model(Lewis et al., 2019). Our implementation followed the parameters of BART-LARGE given in the original paper, with 24 layers, 16 heads and 1024 hidden dimensions. The training process on our meta training and validation dataset took 1 day on 2 Nvidia-3090 GPUs.

3.3 Results

We can discover in Table 2 that pre-trained language model prompting methods significantly outperformed meta-learning baseline and supervised learning baseline. Pre-trained models with manually crafted prompts obtained performance improvement of 46.16, 71.37 and 10.79 in terms of **BLEU-4**, **METEOR** and **NIST** score. Furthermore, our model AUTOGRAPHEX outperformed manual prompting methods by 9.3%, 2.6% and 30.8% in terms of **BLEU-4**, **METEOR** and **NIST** score. These results showed that AUTOGRAPHEX improves the quality of the generated definition to a large scale.

4 Conclusion

In this work, we tackled the challenge of biomedical definition generation through introducing a zero-shot definition generation dataset, GRAPHINE-ZERO, and a pre-trained language model with automatic prompting mechanism, AUTOGRAPHEX. We examined the performance of AUTOGRAPHEX on GRAPHINE-ZERO and experimental results showed that our method significantly outperformed baseline methods. Experimental results also demonstrated that fully supervised learning methods may fail to perform well in small data regimes, which could be a more realistic scenario. AUTOGRAPHEX can effectively utilize cross domain similarities in definition sentence structure and recognize text descriptions on various biomedical subdisciplines, suggesting that it would be more applicable in real definition generation tasks which suffer from the problem of scarce labelled data.

References

Darrin Baines and RJ Elliott. 2020. Defining misinformation, disinformation and malinformation: An urgent need for clarity during the covid-19 infodemic. *Discussion Papers*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *EMNLP-IJCNLP*, pages 3615–3620.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *NIPS*.

James J Cimino, Paul D Clayton, George Hripcsak, and Stephen B Johnson. 1994. Knowledge-based approaches to the maintenance of a large controlled medical terminology. *Journal of the American Medical Informatics Association, 1(1):35–50*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *HLT 2002*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. *ICML 2017*.

Tianyu Gao, Adam Fischz, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. *ACL 2021*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

Kexin Huang and Marinka Zitnik. 2020. Graph meta learning via local subgraphs. *NeurIPS 2020*.

Simon Jupp, Tony Burdett, Catherine Leroy, and Helen E Parkinson. 2015. A new ontology lookup service at embl-ebi. In *SWAT4LS*, pages 118–119.

Hae Beom Lee, Hayeon Lee, Donghyun Na, Saehoon Kim, Minseop Park, Eunho Yang, and Sung Ju Hwang. 2020a. Learning to balance: Bayesian meta-learning for imbalanced and out-of-distribution tasks. In *International Conference on Learning Representations*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020b. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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

Zhaojiang Lin, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. *ACL 2019*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ArXiv*, abs/2107.13586.

Zequn Liu, Shukai Wang, Yiyang Gu, Ruiyi Zhang, Ming Zhang, and Sheng Wang. 2021b. Graphine: A dataset for graph-aware terminology definition generation. *EMNLP 2021*.

Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L Rubin, Margaret-Anne Storey, Christopher G Chute, et al. 2009. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, pages W170–W173.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *ACL 2002*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Technical report*.

Timo Schick and Hinrich Schütze. 2021. Few-shot text generation with natural language instructions. In *EMNLP*.

Ronald Seoh, Ian Birle, Mrinal Tak, Haw-Shiuan Chang, Brian Pinette, and Alfred Hough. 2021. Open aspect target sentiment classification with natural language prompts. *EMNLP 2021*.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *EMNLP 2020*.

Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, and Christopher J Mungall. 2007. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*.

Yiping Song, Zequn Liu, Wei Bi, Rui Yan, and Ming Zhang. 2020. Learning to customize model structures for few-shot dialogue generation tasks. *ACL 2020*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NIPS 2017*.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3914–3923, Hong Kong, China. Association for Computational Linguistics.