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 employing randomly generated 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
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 sub-words 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 $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$ denote N graphs. For each graph $G = \{V, E, X, Y, d\}$, $V$ is a set of nodes, $E$ is a set of edges, $d$ is the description sentence for the graph $G$. Each node is associated with a terminology $x_i \in X$ and a definition $y_i \in Y$. 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 $G \in G_{test}$.

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...
The idea of AUTOGRAPHEX 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:

\[ V_{\text{cand}} = \max_{\omega_C} \langle \omega_C^T \nabla_{\omega_C} \log(p(X_{\text{prompt}})) \rangle \]  

where \( \omega \) is the embedding of the candidate token, \( \omega_t \) 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 \( X_{\text{prompt}} \) 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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>( X_{\text{prompt}} )</th>
<th>( Y_{\text{prompt}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>[Term] is in [Desc]. Concat([Term] is a [Parent-1],...,[Term] is a [Parent-n]). Concat([Child-1] is a [Term],...[Child-n] is a [Term]). The definition of [Term] is [Def]</td>
<td>[Def]</td>
</tr>
<tr>
<td>AutoGraphex</td>
<td>[Term] [T] [T] [T] [Desc]. Concat([Term] [T] [T] [Parent-1],...,[Term] [T] [Parent-n]). Concat([Child-1] [T] [T] [Term]). [T] [T] [T] [Term]</td>
<td>[Def]</td>
</tr>
</tbody>
</table>

Table 1: Manually crafted prompt templates and the prompt templates used by AUTOGRAPHEX.
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 sub-disciplines 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 BLEU-1, BLEU-2, BLEU-3, BLEU-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 auto-regressive 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 sub-disciplines, suggesting that it would be more applicable in real definition generation tasks which suffer from the problem of scarce labelled data.
References


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


