Prompt-based Zero-shot Text Classification with Conceptual Knowledge

Yuqi Wang[1,3], Wei Wang[1], Qi Chen[1], Kaizhu Huang[2], Anh Nguyen[3], Suparna De[4]
[1]Xi'an Jiaotong Liverpool University, China
[2]Duke Kunshan University, China
[3]University of Liverpool, United Kingdom
[4]University of Surrey, United Kingdom
yuqi.wang17@student.xjtlu.edu.cn, {wei.wang03,qi.chen02}@xjtlu.edu.cn,
kaizhu.huang@dukekunshan.edu.cn, anh.nguyen@liverpool.ac.uk, s.de@surrey.ac.uk

Abstract

In recent years, pre-trained language models have garnered significant attention due to their effectiveness, which stems from the rich knowledge acquired during pre-training. To mitigate the inconsistency issues between pre-training tasks and downstream tasks and to facilitate the resolution of language-related issues, prompt-based approaches have been introduced, which are particularly useful in low-resource scenarios. However, existing approaches mostly rely on verbalizers to translate the predicted vocabulary to task-specific labels. The major limitations of this approach are the ignorance of potentially relevant domain-specific words and being biased by the pre-training data. To address these limitations, we propose a framework that incorporates conceptual knowledge for text classification in the extreme zero-shot setting. The framework includes prompt-based keyword extraction, weight assignment to each prompt keyword, and final representation estimation in the knowledge graph embedding space. We evaluated the method on four widely-used datasets for sentiment analysis and topic detection, demonstrating that it consistently outperforms recently-developed prompt-based approaches in the same experimental settings.

1 Introduction

Numerous studies have achieved great success in applying supervised natural language processing (NLP) techniques to address a plethora of NLP applications, including text classification (Dong et al., 2019), natural language inference (Wang et al., 2020) and neural machine translation (Mi et al., 2016). However, achieving high accuracy with deep learning models for textual data analysis necessarily requires a large amount of manually annotated samples, which is both time-consuming and labour-intensive.

To address the issues in low-resource settings, considerable attention has been paid to the pre-trained language models (PLMs), such as GPT-3 (Brown et al., 2020), BERT (Devlin et al., 2019), and Roberta (Liu et al., 2019), due to their superior performances on knowledge transfer. The model pre-training stage typically involves language modelling tasks, i.e., word prediction based on the context of the input. Extensive investigations, e.g., knowledge probing, on PLMs show that they have a certain capacity to store both linguistic and relational knowledge from large-scale corpora of general domain data (Petroni et al., 2019).

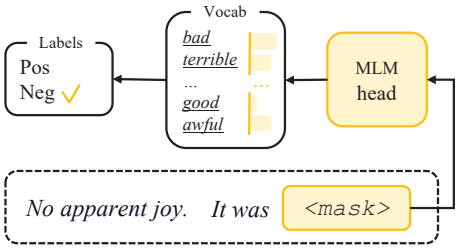

Figure 1: An example of prompt-based text classification for the binary sentiment analysis task.

In recent years, the paradigm of NLP has been shifted from "pre-train and fine-tune" to "pre-train and prompt" (Liu et al., 2023), to fully exploit these PLMs in a gradient-free manner and effectively mitigate the gap between pre-training tasks and downstream tasks for the extreme zero-shot scenario (Yin et al., 2019). Specifically, in the prompt-based approaches (Schick and Schütze, 2021; Min et al., 2022; Gao et al., 2021a), each sample in NLP tasks can be wrapped into cloze-style questions with their corresponding templates, prompting the PLMs to generate the targeted output to solve the problem. For example, in a binary sentiment analysis task (shown in Figure 1), the text "*no apparent joy*" is transformed to the prompt-augmented input "*no apparent joy. It was <mask>.*", where the *<mask>* is a special token to be predicted by the PLMs. This text will then be labelled as positive or negative according to the predicted words. Most existing works utilize a verbalizer to provide the translations

from the predicted vocabulary to the label space in a specific task (Schick and Schütze, 2021). However, these approaches are subject to two significant limitations: (i) by only considering a limited set of pre-defined label words filled in the masked position, some potentially relevant or useful words in the certain domain could be ignored, hindering the model's capacity to generalize; and (ii) the pre-training data of PLMs may contain biases that are reflected in the model's predictions on downstream tasks (Zhao et al., 2021). Some works propose calibration strategies to adjust the distribution of prior probabilities (Hu et al., 2022), which requires access to a large amount of data in specific datasets for true estimation.

In this work, we propose a framework to perform prompt-based zero-shot text classification with conceptual knowledge and overcome the above limitations. The proposed framework includes prompt-based keyword extraction, weight assignment to each keyword in the meaningful semantic space, and final representation estimation. Specifically, in the weight assignment component, by leveraging the contextual relationships captured by SimCSE (Gao et al., 2021b), a powerful contrastive learning model, we refine the probabilities of each keyword being filled in the masked position from the language prompt to mitigate the bias. Additionally, in the final representation, we integrate structured factual data provided by the knowledge graphs (KGs) to include a wider range of semantic relationships between entities in a given domain. By combining their strengths, the proposed framework enables more informed predictions and a richer understanding of the underlying domain. In the experiment, we strictly follow the "label-fully-unseen" setting proposed by Yin et al. (2019) for evaluation. We employ four widely-used text classification datasets and compare the proposed framework with several recently-developed prompt-based approaches under the same experimental settings. The result indicates that our proposed framework brings significant improvement to the model performance.

2 Related Works

Language prompt has been introduced to elicit knowledge from PLMs to solve different NLP tasks, which was inspired by a series of works related to prompt-based approaches, including GPT-3 (Brown et al., 2020) and PET (Schick and Schütze, 2021). However, one issue under the zero-shot setting identified by Chen et al. (2022) is the lack of domain adaptation. They performed prompt-aware continual pre-training based on adaptively retrieved data for better performance on text classification tasks. To widen the coverage of label words, Hu et al. (2022) incorporated external knowledge bases for the verbalizer construction, which greatly improved the stability.

The above-mentioned works used hand-crafted prompt templates, particularly designed by humans for various NLP tasks. While they are carefully constructed, the process requires a considerable amount of human effort. Several automatic prompting techniques were introduced to automatically select a prompt based on the input provided to the PLMs. Gao et al. (2021a) suggested to employ a pre-trained text-to-text transformer, T5 (Raffel et al., 2020), for candidate template generation. The best language prompt can be derived after the evaluation of each candidate template. Shin et al. (2020) proposed a gradient-based approach to search for a set of impactful tokens as the prompts that can cause significant changes in the model's output. Nevertheless, the quality of the automatically generated prompt usually cannot be guaranteed, and this approach lacks sufficient interpretability. Besides discrete prompts, research such as (Li and Liang, 2021) and (Gu et al., 2022) presented continuous prompts as prefixes to the input, which are continuous vectors that can be learned based on patterns and structures from the data. This approach avoids the hassle of explicit prompt design while it introduces a large number of new parameters to be optimized.

3 Methodology

We propose a prompt-based approach to tackle the zero-shot text classification problem. The overall framework is shown in Figure 2. We first extract the keywords to summarize the input text with the prompt-based approach. Then, we assign weights to these keywords based on their semantic relevance to the overall meaning of the text. The weighted embeddings of all extracted keywords in the knowledge graph (KG) embedding space are aggregated to produce the final representation of the input text. Finally, we determine if the text is related to a label in the KG according to their cosine similarity. In the following subsections, we describe the task definition in the extreme zero-shot setting, prompt-based keyword extraction, weight

assignment and final representation estimation in the constructed KG embedding space.

3.1 Task Definition

Given n textual inputs $\mathcal{X} = \{x_1, x_2, ..., x_n\}$, the aim of the text classification task is to assign each input x a label y from a fixed label set containing m labels, i.e., $\mathcal{Y} = \{y_1, y_2, ..., y_m\}$. Unlike the label-partially-unseen zero-shot text classification, where a part of labelled data is available for model training or fine-tuning on a specific domain, in this work, all samples are unseen, and only the label names from the label set $\mathcal{Y}$ can be accessed in advance. In order to achieve this goal, it is essential to ensure that the aspect being described in the input text and the meanings of the labels are comprehensible to the framework (Yin et al., 2019).

3.2 Prompt-based Keyword Extraction

To remove noise and preserve the most relevant information, keyword extraction from the input text can summarize its main content and identify the most important concepts. The meaning of an expression, particularly its implicit meaning, can often be inferred from the context in which it is used. Therefore, we first employ a contextualized pre-trained masked language model, denoted as $\mathcal{M}$, for prompt-based keyword extraction. This model has an MLM head on top of the transformer-based architecture, and consequently, it reduces the text classification to the MLM problem with a task-specific template t, which is either added at the beginning or the end of the original input to form a prompt-augmented input. The template includes a mask token <*mask*>, and the probability of each word v from vocabulary $\mathcal{V}$ being filled in this position can be predicted by $\mathcal{M}$. The most likely words generated in this manner are somewhat relevant to the input context, as the model integrates contextual information to make predictions. We then construct a keyword set for x, namely, $\mathcal{V}^x$, i.e.,

$$\mathcal{V}^x = \underset{v \in \mathcal{V}}{\text{top}\, K} \left[P_{\mathcal{M}} \left(\text{<*mask*>} = v | [x; t] \right) \right] \quad (1)$$

where $[x; t]$ is the prompt-augmented input for x. $P_{\mathcal{M}}(|\cdot)$ is the conditional probability generated by the MLM head of $\mathcal{M}$. According to the observations by Meng et al. (2020), the top 50 probable words usually well represent the mask. Hence, we set the parameter K to 50.

3.3 Weight Assignment

To estimate the text representation for the input, each word in the $\mathcal{V}^x$ should be associated with a weight, indicating relevance and importance to the original textual input. Directly using the probability output by the MLM head could be one possible solution. However, the masked language model may produce a biased probability distribution over the keyword set.

To address this issue, we utilize SimCSE (Gao et al., 2021b), a Siamese network for simple contrastive learning, to assign weights to each word. SimCSE employs entailments and contractions from natural language inference (NLI) datasets as supervised signals. In contrastive loss, the premise and entailment hypothesis are considered positive pairs, while in-batch negatives and contradiction hypothesis are treated as negative pairs. This approach helps align semantically similar sentence embeddings while separating contradicted/unrelated sentence embeddings.

We use the encoding function for SimCSE $f_\theta(\cdot)$, parametrized by θ, to transform both the original input x and a template in which the mask token has been replaced by the k-th word in $\mathcal{V}^x$, denoted as $\tilde{t}_k$, into a meaningful semantic space. We then assign the weight w_i to the i-th word in $\mathcal{V}^x$ based on the similarity between $\tilde{t}_i$ and x, i.e.

$$w_i = \frac{e^{\text{sim}\left(f_\theta(x), f_\theta(\tilde{t}_i)\right)}}{\sum_{k=1}^{K} e^{\text{sim}\left(f_\theta(x), f_\theta(\tilde{t}_k)\right)}} \quad (2)$$

where $\text{sim}(\cdot)$ is the cosine similarity function.

3.4 Final Representation in Knowledge Graph Embedding Space

As for the extreme zero-shot scenario in our work, ideally, each label y in the label set $\mathcal{Y}$ should be equipped with auxiliary information, e.g., a textual description and hand-engineered attributes. Nevertheless, such information available for a particular task is usually limited and may not provide a precise description of the label. Fortunately, there is a source of external knowledge that can be applied with little human effort – KGs. ConceptNet (Speer et al., 2017) is a type of KG that organizes and represents linked open data regarding real-world entities and their relations, offering rich structured knowledge at the conceptual level for the labels.

To leverage the knowledge from the ConceptNet, a process called retrofitting (Faruqui et al., 2015) is used to refine the pre-trained distributional word

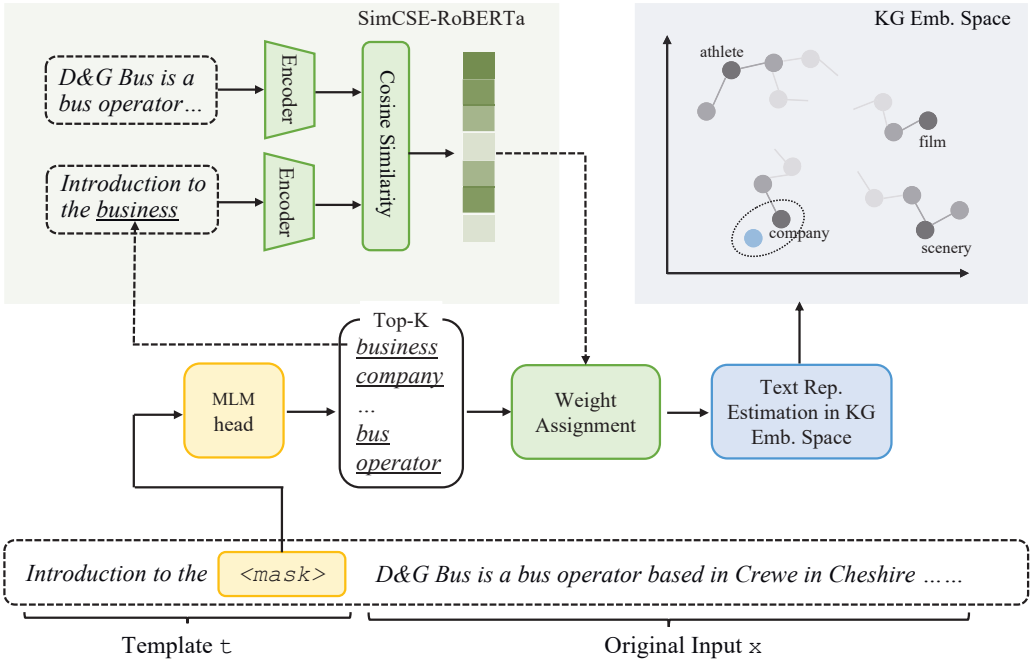

Figure 2: Overall framework of our proposed method

embeddings. The idea is to bring the embeddings of connected entities in the KG closer while maintaining the original distributional ontology (Speer et al., 2017).

The following objective function is minimized to construct the KG embedding space based on the entity set, denoted as $\mathcal{V}^{\text{ent}}$:

$$\sum_{v_i \in \mathcal{V}^{\text{ent}}} \left[\sum_{(v_i, r, v_j) \in \mathcal{E}} \lambda_r \left(\mathbf{v}_i - \mathbf{v}_j\right)^2 + \eta_i \left(\mathbf{v}_i - \hat{\mathbf{v}}_i\right)^2 \right] \tag{3}$$

where $\mathcal{E}$ is the triplet set of the KG, consisting of two entities v_i and v_j linked by their relation r, i.e., (v_i, r, v_j), and λ_r is the corresponding weight for r. $\mathbf{v_i}$ is the updated KG graph embedding for the entity v_i. $\hat{\mathbf{v}}_i$ stands for the original word embedding of v_i and η_i controls the associative strength between $\hat{v}_i$ and v_i. For simplicity, we applied the alignment by the name to align the entity in $\mathcal{V}^{\text{ent}}$ with a word in $\mathcal{V}$.

To estimate the final representation in the KG embedding space for input text x, we integrate the conceptual representation of each keyword v_i in $\mathcal{V}^x$ based on semantic relevance between v_i and x. Our assumption for the multi-class classification task is that the content of input text should remain within its desired label and not be relevant to any other labels in the label set. Therefore, the label with the highest similarity to this representation, among all labels in $\mathcal{Y}$, is then selected as the predicted label, denoted by $\hat{y}$, i.e.

$$\hat{y} = \underset{y \in \mathcal{Y}}{\text{argmax}} \left[\text{sim} \left(\mathbf{v}_y, \sum_{v_i \in \mathcal{V}^x} w_i \mathbf{v}_i \right) \right] \tag{4}$$

where $\mathbf{v}_y$ is the label embedding for y in the KG embedding space.

4 Preliminary Results

4.1 Datasets

We conducted experiments on four commonly used text classification datasets, including two sentiment analysis datasets (SST-2 (Socher et al., 2013) and Yelp-polarity (Zhang et al., 2015)) and two topic detection datasets (AG's News (Zhang et al., 2015) and DBPedia (Lehmann et al., 2015)). We adopted the prompt templates from (Chen et al., 2022) for better comparison. For each dataset, we evaluated our method on different templates and reported their average accuracy along with standard deviation. The statistics and example prompt templates of these datasets are listed in Table 1.

4.2 Setup

For the prompt-based keywords extraction and weight assignment, we made use of roberta-large

Datasets	#Samples	#Classes	Type	Example Prompt
SST-2	1,821	2	Sentiment	All in all, it was <*mask*>
Yelp-polarity	38,000	2	Sentiment	All in all, it was <*mask*>
AG's News	7,600	4	Topic	This topic is about <*mask*>
DBPedia	70,000	14	Topic	Introduction to the <*mask*>

Table 1: Statistics of datasets and example prompt templates used in our work.

models with transformers[1] and simcse[2] libraries. We used the latest version of ConceptNet (5.7) [3] for KG embedding space construction.

We implemented our method with PyTorch 1.5.0 and Python 3.6 on IBM Power 9 architecture. The inference process was accelerated on an NVIDIA Tesla V100 Volta GPU card with 32GB of graphics RAM.

4.3 Main Results

We compared the results with those produced by several prompt-based methods for text classification introduced recently, which share the same extreme zero-shot setting. The main results on the four datasets are shown in Table 2. Channel is the noisy channel approach based on GPT-2 proposed by Min et al. (2022). GPT-3 refers to the work of Zhao et al. (2021) that calibrated the probability distribution with a content-free input. The results of applying Roberta for prompt-based text classification were reported by Chen et al. (2022). AdaPrompt (Chen et al., 2022) refers to the method that adaptively retrieves data from large-scale corpora for continual pre-training, and iAdaPrompt is the process of iterative adaption.

It is clear that the proposed method outperformed the baselines on all datasets, providing a performance gain of 13.88% and 5.31% on Yelp-polarity and AG's News datasets, respectively. Another notable observation from the main results is that our method has significantly lower standard deviations in comparison with Roberta, AdaPrompt and iAdaPrompt, suggesting that it is more stable when using different prompt templates for text classification.

4.4 Ablation Study

We also carried out ablation experiments to explore the effectiveness of weight assignment and KG embedding space construction in the proposed framework. The result of the study is shown in Table 3.

Instead of assigning weights to each keyword based on their importance and relevance as explained in Section 3.3, we directly utilized probabilities of masked token output by the MLM head. This resulted in a slight decrease in performance, with an average accuracy drop of 0.87%. Then, we replaced the KG embeddings for text representation estimation with another semantically consistent embedding, GloVe (Pennington et al., 2014), which is solely based on the word co-occurrence in the pre-training corpus. We observe significant decreases in accuracy on AG's News and DBPedia datasets by 19.3% and 14.4%, respectively. This indicates that, compared with distributional semantic embedding space, incorporating knowledge to construct KG embedding space can greatly enhance the performance of text classification, especially on topic detection datasets.

4.5 Visualization

To further understand the weight assignment, we provided the visualization (shown in Figure 3) of each extracted keyword from examples in topic detection datasets. We arranged these words in descending order of probabilities output by the MLM head. The colour depth denotes the importance of each word according to the given context. As can be seen, many of the most significant keywords (indicated as dark colours) were correctly highlighted. For example, "*rocket*", "*space*" and "*launch*" in AG's News example; "store", "company" and "business" in DBPedia example. We also observed that some less related or wrongly-predicted words could be detected by the model. For example, the DBPedia example mainly describes a game company, even though the words like "*author*" and "*blog*" predicted by the MLM head are at the top of the list, they were assigned with low weights (indicated as light colours) in the weight assignment process, which makes reasonable amendments to the prompt-based keywords

[1] https://huggingface.co/transformers

[2] https://pypi.org/project/simcse/

[3] https://github.com/commonsense/conceptnet-numberbatch

Models	SST-2	Yelp-polarity	AG's News	DBPedia
Channel (Min et al., 2022)	77.10 (N/A)	–	61.80 (N/A)	51.40 (N/A)
GPT-3 (Zhao et al., 2021)	75.80 (0.00)	–	73.90 (0.00)	59.70 (0.00)
Roberta (Chen et al., 2022)	64.56 (16.77)	72.63 (6.34)	69.52 (6.96)	56.32 (0.49)
AdaPrompt (Chen et al., 2022)	75.92 (17.36)	75.09 (17.57)	76.55 (7.28)	70.95 (8.80)
iAdaPrompt (Chen et al., 2022)	77.18 (17.96)	75.81 (18.05)	74.28 (9.00)	73.01 (6.70)
Ours	**80.62 (10.08)**	**89.69 (2.81)**	**81.86 (0.75)**	**73.77 (2.55)**

Table 2: Main results on four commonly-used datasets. We report the average accuracy on different templates and the corresponding standard deviation, which is indicated in brackets.

	SST-2	Yelp-polarity	AG's News	DBPedia
Ours	80.62 (10.08)	89.69 (2.81)	81.86 (0.75)	73.77 (2.55)
-WA	79.42 (10.91)	88.82 (3.08)	81.65 (0.79)	72.59 (2.86)
△	-1.20	-0.87	-0.21	-1.18
-KG	77.58 (10.27)	86.61(4.03)	62.35 (16.16)	58.19 (6.49)
△	-1.84	-2.21	-19.3	-14.4

Table 3: Ablation study. "-WA" means that we directly use the output probability from the MLM head, and "-KG" means that, for final representation estimation, we employ the distributional semantic embedding space rather than KG embedding space.

extraction.

We also demonstrated an example of KG embeddings to show how knowledge integration can help language understanding in Figure 4. We randomly selected a number of generated keywords from samples labelled as "sport","politics", "business" and "technology", and utilized the visualization tool, t-SNE[4], to visualize their corresponding entity embeddings in the two-dimensional space. The colour of each point in the figure indicates the label of the sample from which the keywords were generated. It is observable that entity embeddings assigned to different labels are well distributed across the KG embedding space, indicating that knowledge integration can help capture diverse conceptual aspects of the entities. On the contrary, the embeddings assigned to the same label are well clustered, suggesting that entities with similar properties are mapped closely together in the KG embedding space.

5 Conclusion

We proposed a prompt-based framework to tackle the text classification problem in the extreme zero-shot setting. We exploited the PLM to extract keywords from input, assigned their weights in the meaningful semantic space and incorporated conceptual knowledge from ConceptNet to estimate the final representation. Evaluation results showed that the method reduced the biases of the MLM head and generalized well on two topic detection and two sentiment analysis datasets, outperforming several recently-developed prompt-based approaches.

Limitations

The current work has several limitations that warrant further investigation. Firstly, due to time constraints, we did not conduct experiments using the proposed framework on few-shot settings or a more challenging multi-label classification task. Secondly, our ablation study in Section 4.4 showed that the framework with the weight assignment resulted in only a marginal improvement in performance, suggesting that SimCSE may not be the most effective method for addressing prediction bias. Therefore, future work will explore alternative modeling approaches for bias reduction. Thirdly, in Section 4.5, we noticed that several irrelevant words are also generated as keywords with the language prompt, which may negatively impact the final representation. To address this issue, a better solution, such as keyword filtering, should be considered to improve the current framework. Lastly, we treated each word as a single atomic entity in the KG embedding space, regardless of its possible different senses or meanings. A more careful treatment of word meanings is necessary to handle the problem of polysemy.

[4]https://lvdmaaten.github.io/tsne/

The Race is On: Second Private Team Sets Launch Date for Human Spaceflight (SPACE.com) SPACE.com - TORONTO, Canada -- A second\\team of rocketeers competing for the #36;10 million Ansari X Prize, a contest for\\privately funded suborbital space flight, has officially announced the first\\launch date for its manned rocket.

space	news	science	rocket	space	launch	nasa	space	commercial	aerospace
technology	featured	news	human	exploration	competition	military	innovation	entertainment	international
business	earth	events	engineering	news	personal	education	progress	miscellaneous	sports
mars	aviation	enterprise	discovery	challenges	research	games	transportation	news	robotics
tech	bold	planetary	humans	lunar	rockets	astro	physics	ideas	flight

(a) AG's News example

The GOAT Store (Games Of All Type Store) LLC is one of the largest retro gaming online stores and an Independent Video Game Publishing Label. Additionally, they are one of the primary sponsors for Midwest Gaming Classic.

company	website	sponsor	site	company	site	game	business	store	publisher
show	website	author	team	sponsor	community	blog	podcast	business	group
store	brand	shop	competition	label	games	owner	manufacturer	series	players
publication	franchise	club	game	campaign	goods	blog	team	industry	firm
vendor	league	corporation	partnership	scene	contest	app	organization	promotion	developer

(b) DBPedia example

Figure 3: Weight visualization examples from two topic detection datasets. The Byte-Pair Encoding (BPE) algorithm for the Roberta model may generate words that have their first letters capitalized or a special symbol added as the prefix. After the generation, we replace them with the names of the entities that they actually refer to in the KG. Therefore, there are several duplicates in the keyword set.

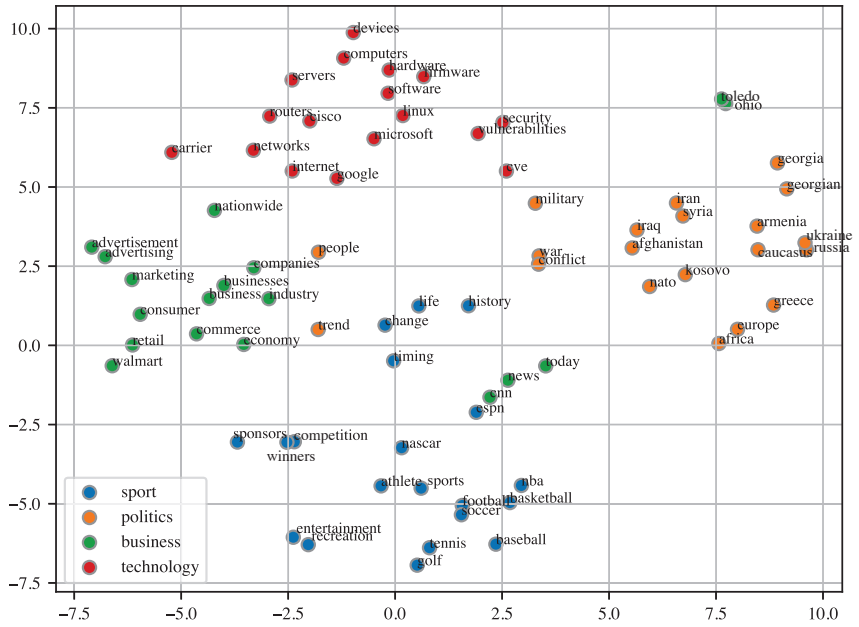

Figure 4: KG embedding visualization. We randomly select several generated keywords from samples labelled as "sport", "politics", "business" and "technology", and utilize the visualization tool, t-SNE, to visualize their corresponding entity embeddings in the two-dimensional space. The colour of each point indicates the label of the sample from which the keyword was generated.

Acknowledgement

We express our sincere gratitude to the matched mentor in the mentoring program, as well as the anonymous reviewers, for their valuable and constructive feedback. Furthermore, we would like to acknowledge the financial support provided by the Postgraduate Research Scholarship (PGRS) at Xi'an Jiaotong-Liverpool University (contract number PGRS2006013). Additionally, this research has received partial funding from the Jiangsu Science and Technology Programme (contract number BK20221260) and the Research Development Fund at Xi'an Jiaotong-Liverpool University (contract number RDF2201132). We are grateful for their support, which has enabled us to carry out this study.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Yulong Chen, Yang Liu, Li Dong, Shuohang Wang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2022. AdaPrompt: Adaptive model training for prompt-based NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6057–6068, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Hang Dong, Wei Wang, Kaizhu Huang, and Frans Coenen. 2019. Joint multi-label attention networks for social text annotation. 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 1348–1354, Minneapolis, Minnesota. Association for Computational Linguistics.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021a. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. PPT: Pre-trained prompt tuning for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8410–8423, Dublin, Ireland. Association for Computational Linguistics.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2022. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2225–2240, Dublin, Ireland. Association for Computational Linguistics.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text classification using label names only: A language

model self-training approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9006–9017, Online. Association for Computational Linguistics.

Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. 2016. Supervised attentions for neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2283–2288.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5316–5330.

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

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473.

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.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics".

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Zikang Wang, Linjing Li, and Daniel Zeng. 2020. Knowledge-enhanced natural language inference based on knowledge graphs. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6498–6508.

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.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.