Query Generation Using GPT-3 for CLIP-Based Word Sense Disambiguation for Image Retrieval

Xiaomeng Pan, Zhousi Chen, Mamoru Komachi*
Tokyo Metropolitan University
{pan-xiaomeng@ed., chen-zhousi@ed., komachi@}tmu.ac.jp

Abstract

In this study, we propose using the GPT-3 as a query generator for the backend of CLIP as an implicit word sense disambiguation (WSD) component for the *SemEval 2023* shared task *Visual Word Sense Disambiguation* (VWSD). We confirmed previous findings — human-like prompts adapted for WSD with quotes benefit both CLIP and GPT-3, whereas plain phrases or poorly templated prompts yield the worst results. Our code is available at https://github.com/pxm427/WSD-for-IR.

1 Introduction

The *SemEval 2023* shared task VWSD[1] combines WSD and Image Retrieval (IR), which aims to select a correct image among ten candidates using a phrase containing ambiguous words. Neural models are likely to be attracted by frequent tokens, labels, and senses of ambiguous words, particularly in limited contexts. We determined that CLIP (Radford et al., 2021) fails to find the correct images using phrases with the ambiguous words of frequent senses even enhanced with contrastive learning on large-scale data.

As shown in Figure 1, given phrase *"Andromeda tree"*, the pretrained CLIP selected incorrect images that focused on either constellation "Andromeda" or part of "tree", neglecting the phrase's meaning entirely. This sample demonstrates ambiguity as a challenge for state-of-the-art neural models. Therefore, we exploited GPT-3 (Brown et al., 2020), a large language model (LLM), for its pretrained knowledge as implicit sense disambiguation and phrase context enrichment for this task.

Prompt engineering boosts model performance and plays a crucial role in applying LLMs to many NLP tasks, which lessens training and testing discrepancies resulting from human-like languages

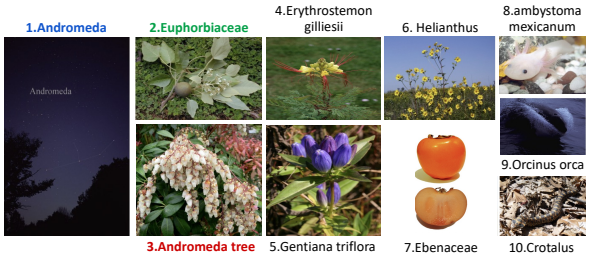

Figure 1: CLIP ranks ten images with respective relevance to the phrase *"Andromeda tree"*, which consists of ambiguous word *"Andromeda"* and limited context *"tree"*. The goal is to select the correct image from ten images of different relevance corresponding to the intended meaning of *"Andromeda tree"*. Note that the first (blue) and second (green) images ranked higher than the correct third image (red).

(Liu et al., 2023). A prompt refers to a text or a set of instructions that guides the model to generate a specific type of response or output. A prompt can be a question, statement, keyword, or sequence of words that provide context and information to the model. Recent research on prompt techniques demonstrates that well-designed prompts spur the potential of neural models without modifying their parameters (Jin et al., 2022). We are curious about how prompts can improve the performance of CLIP on VWSD.

Our main contributions are as follows:

1. We explored different templates for queries and observed their effects on VWSD. The quotes and highlighting ambiguity are effective.

2. We adopted GPT-3 as a key VWSD component to generate queries, which improved the performance in terms of accuracy.

3. We demonstrated that our prompt techniques are effective for finetuning CLIP.

*Now at Hitotsubashi University
[1]https://raganato.github.io/vwsd/

*Proceedings of the The 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 417–422

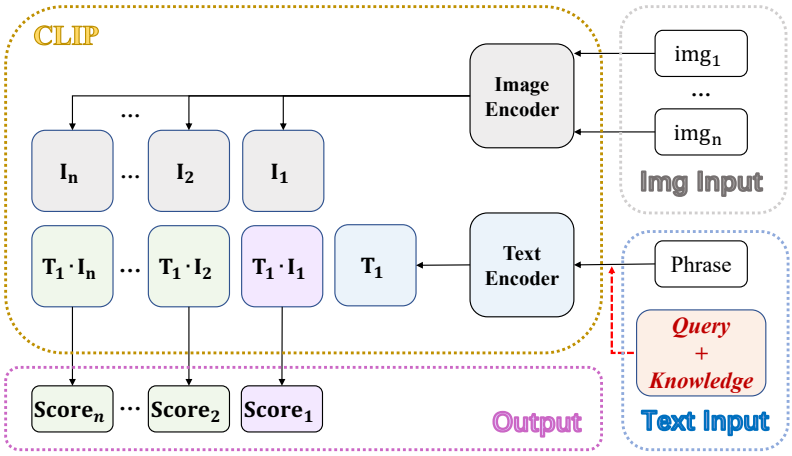

Figure 2: Overview of our method. The left part framed by a brown dotted line is the structure of CLIP, and the right part framed by a blue dotted line and a gray dotted line are our Text and Image Inputs. The bottom part framed by a purple dotted line is our Output.

2 Method

Each VWSD phrase contains an indication of the ambiguous word(s). As illustrated in Figure 2, we first introduce a baseline CLIP that takes text input as either VWSD short phrases or a longer query enhanced with templates for ambiguous words. Then, we leverage GPT-3 to further enrich these queries for CLIP.

CLIP with phrase and queries In a shared feature space, CLIP provides a joint embedding representation for each $(image,\ text)$ pair. The joint embedding representation allows for semantic similarity comparisons between images and text, that is

$$similarity\ score = \text{CLIP}(E_{image}, E_{text}),$$

where E_{image} and E_{text} are the embeddings obtained from its image and text encoders, respectively. We focus on $text$ input for VWSD.

$text$ takes the form of either a single VWSD **phrase** or list of **queries** that bears the phrase and indication of the ambiguous words in the phrase. Table 1 lists our nine templates for creating the queries. Take "*Andromeda tree*" as an example; "*Andromeda*" fits the slot **[ambiguous word(s)]** and "tree" fits **[rest of word(s)]**. Template #1 appears to be logically contradicted with #2. This is because some ambiguous words in the phrase do not fit slot **[ambiguous word(s)]**, but fits slot **[rest of word(s)]**. Moreover, template #3 is for both ambiguous words and rest of words to improve the coverage. These different templates semantically fit different phrases in VWSD, and their performance with CLIP are similar. We select the maximum of the *similarity scores* from all the queries, and we want the image with the highest score.

Query Enrichment with GPT-3 GPT-3 is a powerful language model that can perform various NLP tasks such as language generation, text classification, and question answering. It was designed to improve upon the limitations of previous language models by training a large-scale neural network on massive amounts of text data. This allows GPT-3 to understand the context and generate coherent and contextually relevant responses to text-based inputs, making it useful for a wide range of NLP applications. Additionally, GPT-3 can be finetuned on specific NLP tasks to further enhance its ability to perform various language-related tasks.

As shown in Table 2, we induce additonal **knowledge** from GPT-3 by posing different questions for VWSD phrases which was used to finetune the CLIP: 1) a direct query, 2) a query with double quotes to highlight the phrase, and 3) adding an explicit phrase to separate the phrase ambiguity (i.e., ambiguous word(s)) from others based on 2). These three types are concatenated to the original phrase as a query to finetune CLIP for better performance.

3 Experiments

3.1 Settings

In this study, we used data resources including images and phrases released by SemEval-2023 Task 1. Here, we chose only the English version from

#	Query template for CLIP
1	**[phrase]** is **[ambiguous word(s)]**, but **[rest of word(s)]**.
2	**[phrase]** is really **[ambiguous word(s)]**, but **[rest of word(s)]**.
3	**[phrase]** is not **[ambiguous word(s)]**, but **[rest of word(s)]**.
4	**[phrase]** is not really **[ambiguous word(s)]**, but **[rest of word(s)]**.
5	**[phrase]** is apparently **[ambiguous word(s)]**, but indeed **[rest of word(s)]**.
6	Actually, **[phrase]** is apparently **[ambiguous word(s)]**, but indeed **[rest of word(s)]**.
7	In fact, **[phrase]** is apparently **[ambiguous word(s)]**, but indeed **[rest of word(s)]**.
8	**[phrase]** is not only **[ambiguous word(s)]**, but **[rest of word(s)]**.
9	**[phrase]** is not really **[rest of word(s)]**, but **[ambiguous word(s)]**.

Table 1: Nine query templates for CLIP.

Prompt Type	Question for GPT-3	Answer as Knowledge for CLIP
Direct	What is the *Andromeda tree*?	Andromeda tree is a species of evergreen shrub that belongs to the genus Pieris ...
Double quotes	What is the *"Andromeda tree"*?	The Andromeda tree is a species of flowering evergreen shrub native to ...
Explicit phrase	Instead of *"Andromeda"* and *"tree"*, what is the *"Andromeda tree"*?	The Andromeda Tree is a species of evergreen shrub or small tree native to ...

Table 2: Three types of prompts for inducing GPT-3 knowledge: direct query, double quotes for a phrase, and explicit phrase to separate ambiguous word(s).

Model	Prompt Type	Dev	Test	
CLIP with phrase only	N/A	71.50	58.53	—
CLIP finetuned with nine queries and GPT-3 knowledge	Direct	90.60	56.16	56.80 (ensemble)
		90.20	55.29	
	Double quotes	93.20	65.44	65.87 (ensemble)
		93.00	65.23	
	Explicit phrase	92.20	66.09	65.01 (ensemble)
		91.80	**66.95**	

Table 3: **Dev** and **Test** accuracy results based on data from *SemEval 2023*. To exclude the effects of randomness, we conducted the experiments twice for each prompt type. **Model** represents different versions in our experiments, where the **baseline** is CLIP (phrase). **Prompt Type** indicates the different prompt types used as mentioned in Table 2.

all the language versions (English, Farsi, Italian). We divided the official training data into a *training set (11,869)* and *development set (500)*. Finally, we evaluated our finetuned models on the *test data (463)*, whose contents are different from the training data.

We employed pretrained CLIP as our baseline to calculate the similarity score between the image and text. In the baseline, we used only a phrase as the input of the text component, and the performance is not good. We further finetuned the CLIP to improve the performance by expanding the phrases to queries, and even enriching queries with GPT-3.

For training, we set the batch size to 100 with 10 epochs and used a learning rate of 1e-7. For GPT-3,

Nine Queries	GPT-3 Knowledge	Finetuning	Dev	Test	Better	Worse
			71.50	58.53	0	0
√			72.40	61.12	64	46
√	√		80.40	64.58	90	56
√		√	88.20	61.34	92	74
√	√	√	**92.20**	**66.09**	107	60

Table 4: Accuracy of ablation experiments testing on **Dev** and **Test**. Table 4 presents the number of **Test** samples becoming **Better** and **Worse** after finetuning. Here, only Prompt Type *Explicit Phrase* is used.

we chose *text-davinci-003*, which was considered the most capable GPT-3 model. *text-davinci-003* exhibited a better performance with a higher quality, longer output, and better instruction-following than the other models.

3.2 Results & Analysis

Table 3 lists the results based on the baseline model and finetuned models using different prompt types.

As shown in Table 3, the accuracy of the finetuned models on *Dev* and *Test* performed better than the baseline model. This proves that finetuning can improve the performance of CLIP. All the results on the *Test* were much lower than those on the *Dev*. This may be because the *Dev* data was obtained from the training dataset, which is thematically different from the *Test* dataset.

For different prompt types, the accuracy on the *Dev* varied. The finetuned CLIP adapting prompt type *Direct* had the lowest overall performance with 90.20, and prompt type *Double quotes* had the highest overall performance with 93.20. A speculative reason for this was the lower knowledge quality when selecting the prompt type *Direct* because GPT-3 tended to not consider the phrase entirely when asking directly, thereby generating inaccurate knowledge. For prompt type *Explicit phrase*, it could reach a point of 92.20.

On the *Test*, including the baseline, prompt type *Explicit phrase* exhibited the best performance, which could reach up to 66.95. This indicates that the knowledge generated by GPT-3 was beneficial. Conversely, the performance of prompt type *Direct* was worse than the baseline, which may indicate that poor knowledge can introduce negative effects.

Finally, we conducted ensemble experiments between each prompt type. The results demonstrated an improvement in accuracy for all prompt types except *Explicit phrase*.

3.3 Ablation Study

To investigate the benefit of the effect of queries, knowledge from GPT-3, and finetuning, we conducted ablation experiments. We counted the number of answers that improved or worsened in terms of the change of the gold answer rank. Table 4 shows that samples that improve are increasing.

Query templates. As shown in Table 4, the baseline results were the lowest: 71.50 and 58.53 on *Dev* and *Test*, respectively. This is because the *phrase* was too short to carry meaningful information. Therefore, when creating a sentence including a target *phrase* as a query, more contextual information can be obtained. Consequently, the score increased by 0.9 and 2.59 on *Dev* and *Test*, respectively, compared with the baseline.

Prompt engineering. To better use of the information in context, we have added knowledge from GPT-3 based on the prompts. The score particularly increased to 80.40 on *Dev*, which proves that adding knowledge from GPT-3 improves the performance.

Finetuning. In the finetuning section, we first finetuned CLIP with only *queries*. After finetuning, the accuracy on *Dev* increased by 15.80 compared with *CLIP with queries*, which proves the importance of finetuning. Paying attention to the results on *Test* in CLIP with queries and GPT-3 knowledge is also important. Table 4 shows the score of 64.58 was 3.24 points higher than the result of *CLIP finetuned with queries*. This is partially explained by the knowledge from GPT-3 being partially effective. We further finetuned CLIP with prompts and GPT-3. The best performance reached scores of 92.20 and 66.09 on *Dev* and *Test*, repectively, which illustrates the usefulness of GPT-3.

4 Related Work

4.1 Knowledge generated from LLMs

Unlike *WordNet* (Miller, 1994) and *SemCor* (Miller et al., 1993), recent large-scale language models (LLMs) provide an easy explanation for ambiguous words. In particular, LLMs are well suited for disambiguation tasks. For example, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have a general language understanding ability that has been demonstrated to capture word senses (Coenen et al., 2019).

Recently, Brown et al. (2020) proposed Generative Pretrained Transformer-3 (GPT-3), an LLM trained on a massive amount of data, for various NLP tasks such as dialogue generation (Zheng and Huang, 2021; Lee et al., 2022). It demonstrates an understanding of logical reasoning and external knowledge, which has made it applicable GPT-3 to solving complex problems involving cause-and-effect relationships (Liu et al., 2022). When provied with a proper prompt or asked a human-like question, a pretrained GPT-3 model responds with fluent and relevant text as an answer, which shows passable "logic" and details for disambiguation. The quality of the answer depends on the prompt and question. However, to the best of our knowledge, no research has been conducted on VWSD using LLMs. In this study, we rely on LLM output as external knowledge for VWSD.

4.2 Image Retrieval

Recently, IR has undergone dramatic shifts from approaches handcrafted with global and local descriptors, to convolutional neural networks (He et al., 2016) with adaptive local descriptors, to recent non-convolutional models with one global descriptor, such as a Vision Transformer (Dosovitskiy et al., 2021, ViT). Experimental evaluations (Gkelios et al., 2021) show that ViT achieves competitive results at a low complexity and even finetuning is not required, which makes it an attractive choice as a baseline model for IR.

Recently, researchers began leveraging natural language descriptions in computer vision to improve performance. He and Peng (2017) and Liang et al. (2020) showcased the utilization of natural language descriptions and explanations to enhance the fine-grained visual classification of birds. Radford et al. (2021) presented CLIP in a zero-shot setting, which demonstrated the model's substantial potential for widely-applicable tasks such as IR (Mori et al., 1999).

5 Conclusion and Future Work

We explored the effects of a query on VWSD and used GPT-3 as a key to generate queries. After finetuning CLIP with queries generated from GPT-3, we determined that queries generated by GPT-3 using prompts improved the performance in terms of accuracy.

In the future, we plan to apply some other multimodal models and compare the results with those of existing works. We also intend to adopt GPT-4 [2] to generate knowledge considering both textual and visual cues for VWSD.

References

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. *CoRR*, abs/2005.14165.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda B. Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. *CoRR*, abs/1906.02715.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations*.

Socratis Gkelios, Yiannis S. Boutalis, and Savvas A. Chatzichristofis. 2021. Investigating the Vision Transformer Model for Image Retrieval Tasks. In *17th International Conference on Distributed Computing in Sensor Systems*, pages 367–373.

[2] https://arxiv.org/abs/2303.08774

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Xiangteng He and Yuxin Peng. 2017. Fine-graind Image Classification via Combining Vision and Language. *CoRR*, abs/1704.02792.

Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. 2022. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2763–2775, Dublin, Ireland. Association for Computational Linguistics.

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does GPT-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Weixin Liang, James Zou, and Zhou Yu. 2020. ALICE: Active learning with contrastive natural language explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4380–4391, Online. Association for Computational Linguistics.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. 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):195:1–195:35.

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. *CoRR*, abs/1907.11692.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Yasuhide Mori, Hironobu Takahashi, and Ryu ichi Oka. 1999. Image-to-word transformation based on dividing.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Chujie Zheng and Minlie Huang. 2021. Exploring prompt-based few-shot learning for grounded dialog generation. *CoRR*, abs/2109.06513.