SUT at SemEval-2023 Task 1: Prompt Generation for Visual Word Sense Disambiguation

Omid Ghahroodi[◇*], **Seyed Arshan Dalili**[◇*], **Sahel Mesforoush**[◇], **Ehsaneddin Asgari** [§]
[◇] DH-NLP Lab, Computer Engineering Department, Sharif University of Technology
[§] AI Innovation Center, Data:Lab, Volkswagen AG, Munich, Germany
{omid.ghahroodi98, sahel.mesforosh}@sharif.edu, adalili@ce.sharif.edu, asgari@berkeley.edu

Abstract

Visual Word Sense Disambiguation (V-WSD) identifies the correct visual sense of a multi-sense word in a specific context. This can be challenging as images may need to provide additional context and words may have multiple senses. A proper V-WSD system can benefit applications like image retrieval and captioning. This paper proposes a Prompt Generation approach to solve this challenge. This approach improves the robustness of language-image models like CLIP to contextual ambiguities and helps them better correlate between textual and visual contexts of different senses of words.

1 Introduction

There are different senses associated with polysemous English words (Bevilacqua et al., 2021). The task of determining a specific sense of a word has been one of the most significant challenges in Natural Language Processing. Several studies have been conducted on determining a word's sense based on its sentence. With the emergence of multimodal works, visual and acoustic information can be used to solve a word's ambiguity. These models can be used to expand capability of models in limited textual context. Using a proper Visual Word Sense Disambiguation model can contribute to the advancement of different areas, such as image captioning, image retrieval, and other multimodal problem settings.

Our main approach is to tackle the problem of V-WSD (Raganato et al., 2023) using prompt engineering. We used ChatGPT (OpenAI, 2022) to generate a desired prompt for a given sense. Using CLIP model (Radford et al., 2021) we extract text and image embeddings for the given input and compare their cosine distance. Our model works for English, Farsi and Italian inputs with average MRR of 71.13% ranking 9th in SemEval final results.

Using ChatGPT to expand contextual information increased model's accuracy. However, since ChatGPT usage has limitation on number of requests, the model input is limited.

Our code is publicly available on github[1].

Related Works

There are substantial works on word sense disambiguation in natural language processing (Navigli, 2009). A word sense disambiguation can be done using supervised (Wang and Wang, 2021), (Bevilacqua and Navigli, 2019), (Hadiwinoto et al., 2019), or knowledge-based models (Moro et al., 2014), (Agirre et al., 2014). Knowledge-based works use lexicon graph-based information (such as WordNet) to extract a word's sense. Supervised methods return the sense of a word based on the presented context (Bevilacqua et al., 2021).

The upper bound accuracy for this WSD is 80 percent which is human accuracy on annotated data (Navigli, 2009). State-of-the-art models (Bevilacqua and Navigli, 2020), (Barba et al., 2021) have reached this accuracy. In V-WSD we use visual information beside contextual data in order to improve model accuracy (Barnard and Johnson, 2005), (Su and Jurie, 2011). There are several datasets for V-WSD. Most of these datasets concentrate on verbs (Gella et al., 2016), (Gella et al., 2019). MuWoSeD (Anonymous, 2023) is a general visual word dataset we use to finetune our models.

2 Our System

2.1 Overview

The overview of our approach to tackling the problem of V-WSD is based on the idea of prompt engineering. Prompt Engineering helps the model

[*]The first two authors contributed equally and listed randomly.

[1]https://github.com/language-ml/SemEval-VWSD

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2160–2163

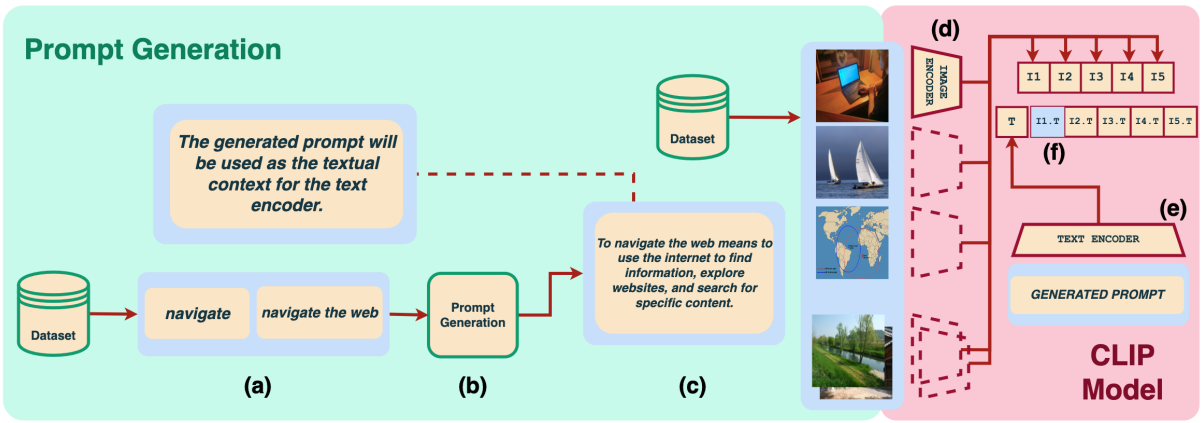

Figure 1: The overall pipeline of the proposed model. **(a):** we select the ambiguous words and their textual contexts from the dataset. **(b):** we feed the Prompt Generation module with words and their corresponding textual contexts to generate prompts for the word. **(c):** we use generated prompts as the textual input of our Multimodal Module. **(d-f):** models are evaluated by computing the similarity between encoded images and the encoded generated prompt.

overcome the disambiguation present in the textual contexts by adding details and clarification in the given prompts to the model to find the corresponding visual context. Large Language Models (LLMs) can help us better understand the context of a sense. As a result, we use ChatGPT as our LLM for generating prompts for each of the senses. Our approach does not need any fine-tuning technique for inference, resulting in more efficiency for the model.

2.2 Prompt Generation Module

Here we detail the Prompt Generation module: we first give the word and corresponding sense to the ChatGPT to generate a desired prompt for that sense of a word. This is done by first transforming the textual context into a template of the form *"What does the [WORD] mean in [CONTEXT]?"*. Then, we get the response from ChatGPT that contains some extra information regarding that word, which can guide the multimodal model to correlate visual and textual contexts better. Furthermore, due to the limit of requests and the large number of words in the test dataset, we batched these words in a request to reduce the number of requests to avoid the ChatGPT limit.

2.3 Multimodal Module

We use the CLIP model (Radford et al., 2021) as our Multimodal Module to have a shared embedding space between texts and images. This module produces an embedding for texts and images such that related texts and images have more cosine similarity than others. CLIP model (Radford et al., 2021) uses a text encoder and image encoder and projects the output vectors to a shared space, calculates the similarity between any two pairs of text embedding and image embedding of a batch of N pair samples into an $N * N$ matrix with cosine similarity, and attempts to increase the value of N elements in the diagonal matrix while decreasing the value of the other $N * N$.

As the CLIP model has been trained on a large amount of data, it should have learned the shared embedding between the image and the text well enough to differentiate between the meanings of ambiguous words and related images.

We provided a visual representation of our approach in Figure 1. We provide the generated prompt and candidate images to the CLIP model, as depicted in **part(d)** and **part(e)** of Figure 1, and based on the cosine similarity between the text em-

Model	Hits@1	MRR
English		
Our model(zero-shot)	.67	.79
Our model(fine-tuned)	**.69**	**.80**
Pre-trained MuWoSeD	.68	.80
Farsi		
Our model(zero-shot)	**.49**	**.64**
Italian		
Our model(zero-shot)	**.53**	**.68**
Average		
Our model(zero-shot)	**.57**	**.71**

Table 1: Results for our models on this task in two settings (Fine-tuning & Zero-shot settings) based on two metrics (Hits@1 & MRR)

bedding and each image embedding, we determine a ranking for the related images, as depicted in **part(f)** of Figure 1.

3 Experimental Setup and Evaluation

For our experiments, each dataset instance has three main parts: the ambiguous word, limited textual context, and ten image candidates, one of which corresponds to the desired visual context, and there were three languages in the test dataset: English, Italian, and Farsi. We first batched the dataset words and their textual contexts and got their corresponding prompts. We then used different backbones of the CLIP model as the Multimodal Module of our system. Additionally, we took a mean pooling of all of these backbones to predict the correct image. For the Farsi and Italian words, we translated them into English and then used the translated text as the input for our Multimodal Module.

We used ResNet-50 and ResNet-101 (He et al., 2016) and ViT (Dosovitskiy et al., 2020) backbones for the CLIP models. We also used fine-tuned models on the MuWoSeD dataset (Anonymous, 2023), a Multimodal Word Sense Disambiguation dataset. We fine-tuned models in 5 epochs on this dataset. We test various models using the same two ranking metrics specified in the task. The first metric is Hits@1. This metric measures the frequency with which the ground truth appears in the top position. Mean Reciprocal Rank (MRR) is our second metric, calculated using the following formula:

$$\frac{1}{N}\sum_{i=1}^{N}(\frac{1}{rank_i})$$

where N refers to the number of test data and $rank_i$ refers to the rank of ground truth in i-th test data.

4 Results

The results for our models are shown in Table 1. For the English language, our model achieved Hits@1 of 0.67 and MRR of 0.79 in zero-shot settings. When fine-tuned on the training dataset of the task, it increased to 0.69 and 0.80 for Hits@1 and MRR, respectively. The pre-trained CLIP on the MuWoSeD dataset got Hits@1 of 0.68 and MRR of 0.80, which shows an increase compared to the zero-shot setting.

For the Farsi language, our model achieved Hits@1 of 0.50 and MRR of 0.64, suggesting that this language is more challenging for the model. Moreover, our model produced Hits@1 of 0.53 and MRR of 0.68 for the Italian language.

Overall, our model achieved Hits@1 of 0.57 and MRR of 0.71. This suggests that our zero-shot approach performs fairly well, given the fact that it is not fine-tuned on a specific dataset and can perform fairly well across different domains. In addition, it consumes less energy since it does not need any data for training. We also evaluated the pre-trained MuWoSeD on this task dataset and observed that the pre-trained model on the same task but with different data increased the accuracy in the zero-shot setting.

5 Discussion and Conclusions

In this paper, we proposed a model to solve the Visual Word Sense Disambiguation challenge using Large Language Models to generate prompts for different textual contexts of words to add extra information to them to make multimodal models select the corresponding image of that context better.

For the English language, we used ChatGPT to generate those prompts and then gave those prompts to different backbones of CLIP models to detect the related image of that sense both in zero-shot and fine-tune settings. It shows that fine-tuning on the training dataset improves the model's performance. Moreover, using a pre-trained model on the MuWoSeD dataset in zero-shot settings leads to better results than the original CLIP model in the zero-shot setting.

We first translated textual contexts for the Farsi and Italian languages and then used the result as the text input of our multimodal models in zero-shot settings. Using a zero-shot setting significantly reduces the consumed energy while maintaining its performance at a fairly good level in different domains.

Future works can include prompt generation in multilingual settings to add extra information to the textual input of the multimodal models.

References

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

Anonymous. 2023. Muwosed: Multimodal word sense

dataset and modeling for visual word sense disambiguation. Anonymous preprint under review.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021. ConSeC: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kobus Barnard and Matthew Johnson. 2005. Word sense disambiguation with pictures. *Artificial Intelligence*, 167(1):13–30. Connecting Language to the World.

Michele Bevilacqua and Roberto Navigli. 2019. Quasi bidirectional encoder representations from transformers for word sense disambiguation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 122–131, Varna, Bulgaria. INCOMA Ltd.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, Online. Association for Computational Linguistics.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Spandana Gella, Desmond Elliott, and Frank Keller. 2019. Cross-lingual visual verb sense disambiguation. 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 1998–2004, Minneapolis, Minnesota. Association for Computational Linguistics.

Spandana Gella, Mirella Lapata, and Frank Keller. 2016. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 182–192, San Diego, California. Association for Computational Linguistics.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. In *Conference on Empirical Methods in Natural Language Processing*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

TB OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

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. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alessandro Raganato, Iacer Calixto, Asahi Ushio, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2023. SemEval-2023 Task 1: Visual Word Sense Disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics.

Yu Su and Frédéric Jurie. 2011. Visual word disambiguation by semantic contexts. In *2011 International Conference on Computer Vision*, pages 311–318.

Ming Wang and Yinglin Wang. 2021. Word sense disambiguation: Towards interactive context exploitation from both word and sense perspectives. 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 5218–5229, Online. Association for Computational Linguistics.