OPI PIB at SemEval-2023 Task 1: A CLIP-based Solution Paired with an Additional Word Context Extension

Małgorzata Grębowiec
National Information Processing Institute, 00-608 Warsaw, Poland
mgrebowiec@opi.org.pl

Abstract

This article presents our solution for SemEval-2023 Task 1: Visual Word Sense Disambiguation. The aim of the task was to select the most suitable from a list of ten images for a given word, extended by a small textual context. Our solution comprises two parts. The first focuses on an attempt to further extend the textual context, based on word definitions contained in WordNet and in Open English WordNet. The second focuses on selecting the most suitable image using the CLIP model with previously developed word context and additional information obtained from the BEiT image classification model. Our solution allowed us to achieve a result of 70.84% on the official test dataset for the English language.

1 Introduction

Visual word sense disambiguation (VWSD) is a classification problem that involves selecting the one image from a set of candidates that most accurately represents the meaning of a target word. This task is important because many words have multiple meanings, and the intended meaning of a word can be ambiguous based on its context. The problem of word ambiguity is a well-established one (Bevilacqua et al., 2021), but VWSD is relatively new and contributions to the subject remain sparse. To popularise this issue, SemEval 2023 Task 1: V-WSD: Visual Word Sense Disambiguation (Raganato et al., 2023) was organised. The organisers prepared the task for three languages: English, Farsi, and Italian. It was possible to participate in the competition for each language separately. Our solution is designed exclusively for English.

This article describes our solution for the competition. We focus chiefly on describing our attempts to extend the context for given ambiguous words. This is crucial for selecting the most appropriate image. The remainder of the publication is organised as follows: Section 2 describes the datasets provided by the organisers for each phase of the competition; Section 3 presents the design of the system in detail; Section 4 offers information on the evaluation metrics used and describes the experiments; Section 5 analyses the results of the experiments; Section 6 contains summarises and presents our conclusions.

2 Dataset description

The datasets were prepared by the organisers as TSV files. Each line represents a sample that comprises a single word, limited textual context, and ten image file names. A separate directory that contains image files was attached to each collection. In addition, for the trial and training set, a so-called gold file was included. It contains information about the best-matching image for each word. After the competition, the organisers released the gold file for the test datasets. Table 1 presents examples of samples from the training and test datasets.

The trial data comprises only sixteen samples and 160 images. The training data comprises 12869 samples and 12999 images. Both collections were prepared for the English language. However, in the training dataset, when single words in another language, such as Polish, Chinese, or Latin were found, the word was replaced by an icon. The test dataset has 463 samples and 8100 images provided in two versions: original size and resized. The collection was prepared for three languages: English, Farsi, and Italian.

3 System Description

The architecture of our system is presented in Figure 1. It consists of two components: the Context Extension Module and the Image Ranking Module. The Context Extension Module is responsible for expanding the textual context of the target words based on their definitions and related terms. The Image Ranking Module focuses on sorting images

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 482–487

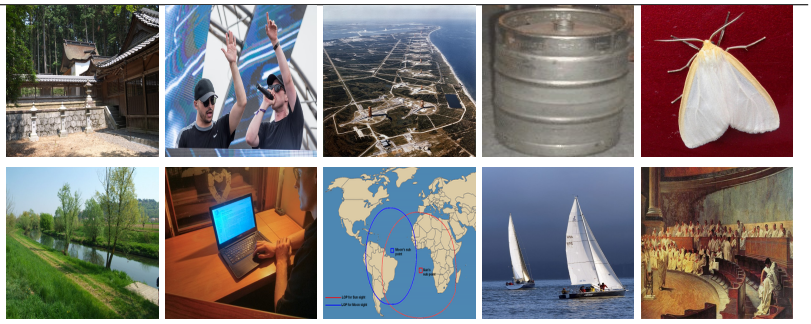
Target word	Context	Image candidates
navigate	navigate the web	
blue	blue mood	

Table 1: Example of samples from the test dataset.

from the best- to the worst-matching context. Detailed descriptions of these modules are presented in the proceeding subsections.

3.1 Context Extension Module

WordNet (Fellbaum, 1998) is a lexical database of the English language in which nouns, verbs, adjectives, and adverbs are grouped into cognitive synonym *synsets*. Open English WordNet (McCrae et al., 2019) is a fork of Princeton WordNet, developed using an open-source methodology, which we used as an alternative source. Each synset is accompanied by a short description, and can be linked to other synsets by means of conceptual-semantic and lexical relations. This structure of WordNet allowed us to assume that there is a chance of finding a connection between the target word w_t and its contextual word w_c. The analysis of this assumption comprises four basic steps.

Recipe

In the first step, the simplest case is studied. We check whether a phrase consisting of a target word w_t and a contextual word w_c occurs in WordNet. If it does, its definition is returned as the final version of the extended context; if it does not, we progress to the next step.

The second step assumes that a target word w_t (or its related terms) occurs in one of the synset's context word (w_c) definitions. For this purpose, both words are lemmatised. All of the synsets for the target word w_t are retrieved. Each synset definition is lemmatised and the number of occurrences of the contextual word w_c and its related terms is counted. By related terms, we mean a list of words that includes hyperonyms, hyponyms, lemmas, meronyms, and holonyms of the word. Next, all definitions for which the number of occurrences was greater than zero are passed for further analysis.

In the third step, all synsets for the contextual word w_c are retrieved. As before, each definition is lemmatised, then the number of occurrences of the target word w_t and its related terms is counted.

The final step uses the BEiT model (Bao et al., 2021). It is an image classifier model that was pretrained in a self-supervised fashion on ImageNet-22, a dataset that contains fourteen million images and 21841 classes. ImageNet is organised according to the WordNet hierarchy, so that its class names correspond to synset lemmas. Based on this, we verify whether, among the definitions and related terms returned from steps 2 and 3, the model returns class names. If it does, the number 1000 is added to the sum of the occurrences from the previous steps. This raises the score for the considered candidate of the extended context. Its value was chosen empirically based on the experiments. The extended context with the highest number of

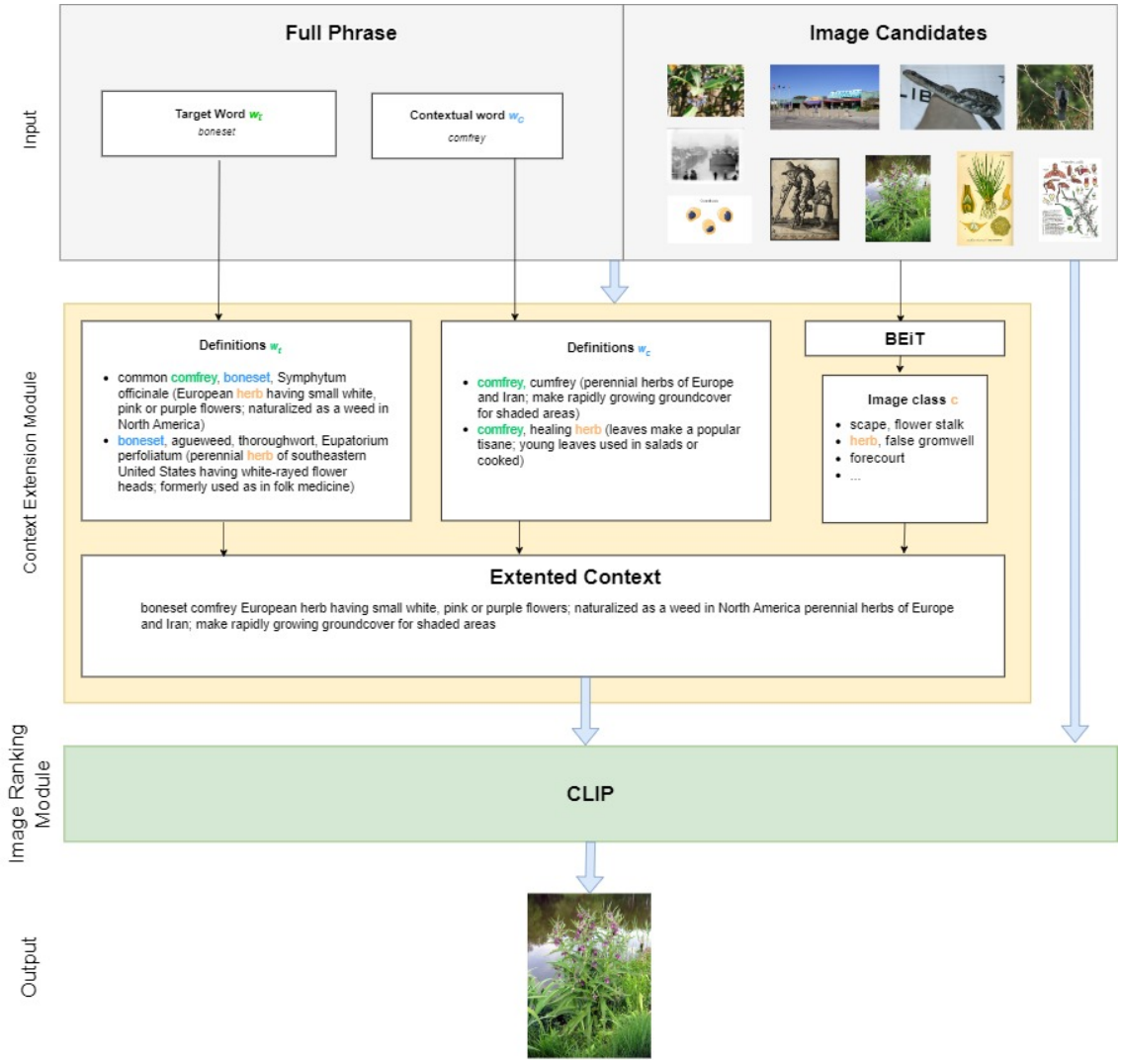


Figure 1: The architecture of our system for the Visual Word Sense Disambiguation task

occurrences progresses to the next module.

3.2 Image Ranking Module

The basic task of the module is to sort the given set of candidate images from best- to the worst-matching of the phrase. For this purpose, the Contrastive Language-Image Pre-Training (CLIP) model was used. CLIP is a deep learning model developed by OpenAI (Radford et al., 2021). It uses a transformer architecture, a type of neural network that has achieved state-of-the-art performance in a wide range of natural language processing and computer vision tasks. The CLIP model uses a variant of the transformer architecture known as the Vision Transformer (ViT), which was developed for image classification tasks. It enables prediction of which of the given images is the most appropriate for the given text. Our solution uses a variant of the CLIP model: CLIP ViT-H/14, developed by LAION (Schuhmann et al., 2022). We used the original model weights, without any fine-tuning. We transform our extended context and image to their vector representations and calculate the similarity between them. We use softmax to obtain the image label probabilities.

4 Experiments

All of our experiments focused on a sub-task for English. To implement the Context Extension Module, we used WordNet v3.1 and Open English WordNet 2021, a derivative solution of the Princeton WordNet developed under open-source methodology. Microsoft's BEiT model[1] was used to classify candidate images. To implement the ranking

[1] https://huggingface.co/microsoft/beit-base-patch16-224-pt22k-ft22k

module, we used the ready-made CLIP model[2] from LAION. For both BEiT and CLIP model we used the original model weights, without any fine-tuning.

4.1 Metrics

The organisers selected mean reciprocal rank (MRR) and a hit rate (HR) of 1 as their official evaluation measures. MRR offers a score that corresponds to the position of the correct answer in the ranking (one for the first position, one-half for the second, one-third for the third, and so on); HR counts only the correctness of the first place in the ranking. Ultimately, the determining measure in the competition was HR.

4.2 Development Results

Table 2 presents the results of the experiments from the trial and test datasets. As a baseline for our solution, we selected the simplest model built from the CLIP model and the original full phrase (target word plus context word) provided from each sample.

For the trial data, the best results were achieved for the context built on WordNet: 92.70% for MRR and 87.50% for HR. Using the BEiT model did not affect the results. This means that no class name that describes the image occurred in the context. For the test data, the best results were achieved for the context built on Open English WordNet: 91.30% for MRR and 86.03% for HR. Since the trial set is very small, the results cannot be considered representative. In further experiments, all of the presented model combinations will be tested.

Model	Trial dataset		Training dataset	
	MRR	**HR**	**MRR**	**HR**
CLIP	88.54	81.25	86.57	79.23
CLIP + WN	**92.70**	**87.50**	91.17	85.84
CLIP + OEWN	89.58	81.25	**91.30**	**86.03**
CLIP + WN + BEiT	**92.70**	**87.50**	91.16	85.83
CLIP + OEWN + BEiT	89.58	81.25	91.29	86.02

Table 2: Results of experiments for the trial and training datasets divided into different context extension methods.

Table 3 presents the distribution for each method of context extension for the training data, divided into WordNet and Open English WordNet. The *Full Phrase* method refers to cases in which a direct composite of the target word and the context word occurs in the dictionary. The method was able to find definitions for 19.33% of all samples in the case of WordNet, and 19.35% in the case of OEWN. Their accuracy for correct image selection was 83.59% and 84.33%, respectively. The *Definitions* method extends the context by matching word definitions based on the number of occurrences of words (and their related terms) from the phrase. The method was able to find definitions for 53.46% of all samples in the case of WordNet, and 39.93% in the case of OEWN. Their accuracy for correct image selection was 86.97% and 85.17%, respectively. The *Image Class* method using the class names returned by the BEiT model extends The *Definitions* approach involves adding a significantly large number to the number of occurrences of words (and their related terms) if the class name (or its related terms) occurs in the phrase. This raises the score for the considered candidate of the extended context. Such instances for WordNet accounted for 27.21% of all samples and 40.72% of Open English WordNet. Their accuracy for correct image selection was 85.17% and 87.65%, respectively.

Extended Context Method	WN		OEWN	
	Acc	**All**	**Acc**	**All**
Full Phrase	83.59	2487	84.33	2490
Definitions	86.97	6880	85.17	5139
Image Class	85.17	3502	87.65	5240

Table 3: Results of experiments for different context extension methods, divided into WordNet and Open English WordNet for the training dataset. *Acc* stands for the accuracy of correct image selection for each method. *All* stands for the number of all samples matched by each method.

4.3 Test Results

Table 4 presents the results achieved for the test dataset. All but the baseline model combinations shown achieved the same 70.84% score for HR. Differences are noticeable only for MRR, whose score we improved by basing the solution on Open English WordNet.

Model	Test dataset	
	MRR	**HR**
CLIP	79.61	68.25
CLIP + WN	81.60	**70.84**
CLIP + OEWN	81.67	**70.84**
CLIP + WN + BEiT	81.62	**70.84**
CLIP + OEWN + BEiT	**81.69**	**70.84**

Table 4: Results of experiments for test datasets divided into different context extension methods.

Table 5 presents the distribution for each method of context expansion for the test dataset, divided

[2] https://huggingface.co/laion/CLIP-ViT-H-14-laion2B-s32B-b79K

into WordNet and Open English WordNet. The *Full Phrase* cases for WordNet and Open English WordNet were 7.34%. Their accuracy for correct image selection was 88.23%. The *Definitions* cases for WordNet accounted for 28.73% of all samples and 69.33% of Open English WordNet. Their accuracy for correct image selection was 74.43% and 66.97%, respectively. The *Image Class* method for WordNet accounted for 63.93% of all samples and 23.33% Open English WordNet. Their accuracy for correct image selection was 67.22% and 76.85%, respectively.

Extended Context	WN		OEWN	
	Acc	**All**	**Acc**	**All**
Full Phrase	88.23	34	88.23	34
Definitions	74.43	133	66.97	321
Image Class	67.22	296	76.85	108

Table 5: Results of experiments for different context extension methods, divided into WordNet and Open English WordNet for the test dataset. *Acc* stands for the accuracy of correct image selection for each method. *All* stands for the number of all samples matched by each method.

Such results allowed us to obtain the 13th place in the English subtask.

5 Error Analysis

The results of the experiments for our system enable us to distinguish three categories of error: errors in matching definitions, errors in classifying images, and subjective judgment in selecting the correct image.

The methods of extending the context with definitions in WordNet or Open English WordNet were based on the assumption that occurrences of keywords (target words, context words, image class names, and their related terms) would appear among them. However, an analysis of the results highlighted that keywords alone were frequently insufficient. A larger dictionary of synonyms would improve the effectiveness of this approach markedly, but unfortunately we didn't have access to such resource.

The BEiT model used to classify the images also did not always return the correct labels; for that reason, its use should be considered as supportive in the selection of the right definitions. Nevertheless (leaving aside the correctness of the image classification), given the number of occurrences of class names in extended contexts (Tables 3 and 5), the use of this approach is justified.

In assessing the correctness of the model, there is a risk that the image that the model should return fails to represent the given phrase clearly. Some phrases do not refer to issues that are easily represented by objects, such as species of animals, plants, or names of objects; they may instead refer to abstract concepts, such as moods (Table 1), which everyone can interpret differently.

6 Conclusion

This article presents our solution for SemEval-2023 Task 1: Visual Word Sense Disambiguation. We first explained the essence of the VWSD task. Next, we presented the system architecture as a proposed solution to the problem. We described the principles of operation of individual modules. We presented an error analysis conducted on the basis of the results of our experiments. We demonstrated that the relatively simple construction of our solution enabled us to achieve a fairly good final result. The error analysis enables us to propose further work on the issue, which should focus primarily on the expansion of the Context Extension Module. Wikipedia resources and further dictionaries of synonyms, such as Roget's 21st Century Thesaurus (Kipfer, 2005) could be used for this purpose.

Our proposed solution is adapted only to English. Taking into account the other languages, it would be worth considering a solution that would address the problem of multilingualism. This is a good opportunity and topic for further research.

References

Hangbo Bao, Li Dong, and Furu Wei. 2021. Beit: BERT pre-training of image transformers. *CoRR*, abs/2106.08254.

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.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

B.A. Kipfer. 2005. *Roget's 21st Century Thesaurus: Updated and Expanded 3rd Edition, in Dictionary Form*. 21st Century Reference. Random House Publishing Group.

John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019. English WordNet 2019 – an open-source WordNet for English. In *Proceedings of the 10th Global Wordnet Conference*, pages 245–252, Wroclaw, Poland. Global Wordnet Association.

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.

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.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.