GPL at SemEval-2023 Task 1: WordNet and CLIP to Disambiguate Images

Shibingfeng Zhang[1,2] and **Shantanu Nath**[1] and **Davide Mazzaccara**[1]
[1]CIMeC, University of Trento [2]Saarland University
{shibingfeng.zhang,shantanu.nath,davide.mazzaccara}@unitn.it

Abstract

Given a word in context, the task of Visual Word Sense Disambiguation consists of selecting the correct image among a set of candidates. To select the correct image, we propose a solution blending text augmentation and multimodal models. Text augmentation leverages the fine-grained semantic annotation from WordNet to get a better representation of the textual component. We then compare this sense-augmented text to the set of image using pretrained multimodal models CLIP and ViLT. Our system has been ranked 16^{th} for the English language, achieving 68.5 points for hit rate and 79.2 for mean reciprocal rank. The code to this project is available on Github[1].

1 Introduction

Many and very common words in language are polysemous, i.e., they have more than one meanings associated with the lexical object (Ježek, 2015). The task of Word Sense Disambiguation (WSD) consists of identify the correct meaning of a word in a specific context, its sense. For example, the word *Andromeda* has three different meanings according to the Cambridge dictionary: the princess from the Greek mythology, the large galaxy and a particular plant with flowers. The objective of WSD is to determine which sense is active in a specific context, such as the sentence "We found the Andromeda growing on the banks of a lagoon of fresh water".

Visual Word Sense Disambiguation (VWSD) is the extension of the traditional WSD task that incorporates visual information in addition to textual context (Gella et al., 2016). In SemEval-2023 Task 1 proposed by Raganato et al. 2023, the objective is to identify the correct image from a pool of candidates that corresponds to the intended meaning of the word in context. For example, the word is *Andromeda*, the context is *Andromeda tree*, and three candidate images are presented in Figure 1. The image located on the left displays a photograph of an Andromeda tree, which is the correct choice, whereas the others images show other objects and are therefore incorrect.

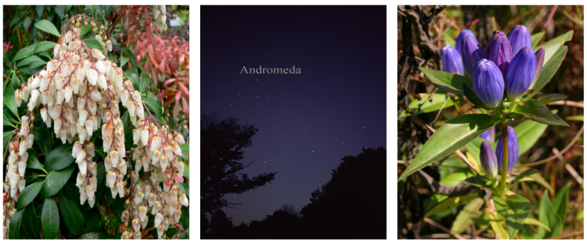


Figure 1: Different images for sense of the words *Andromeda*. The correct image is the one on the left.

We tackle this task with an elaborated text augmentation strategy and the power of pre-trained language-vision models. In particular, our text augmentation strategy employs the fine-grained linguistic resource of WordNet (Miller, 1994) to get better textual representations. Multimodal models are then leveraged to select the correct image in the pool of candidates in a zero-shot. Our system has been ranked 16^{th} for the English language (50 submissions, 125 participants), achieving 68.5 points for hit rate and 79.2 for mean reciprocal rank. We also achieved 37.7 for hit rate and 55.7 for mean reciprocal rank on the Italian dataset.

This report is composed of six sections. Section 2 provides an overview of the history of WSD task and VWSD task. Section 3 outlines our approach to the VWSD task. Section 4 presents the experiment setup. Section 5 presents the experiment settings. The results are presented and evaluated in Section 5. Finally, conclusions are drawn in Section 6.

[1] https://github.com/Zhangshibf/SemEval2023Task1_Visual_Word_Sense_Disambiguation

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1592–1597
July 13-14, 2023 ©2023 Association for Computational Linguistics

2 Related Work

In the traditional Word Sense Disambiguation task (Jurafsky and Martin, 2009), a keyword, the context in which it is utilized, and a collection of potential senses are supplied. The objective of the task is to determine the accurate sense from the set of senses provided, based on the contextual information. Previous studies in the field of WSD mainly leverage knowledge-based methods that relied on manually crafted rules or knowledge resources (Raganato et al., 2017). The Lesk algorithm (Lesk, 1986) was one such approach that select the sense with the highest n-gram overlap with the context of the target word, and has been widely recognized as a robust baseline method for WSD. Some knowledge-based methods also use graph-based algorithms for WSD, such as the approach presented in Agirre and Soroa (2009), which employs random walks over a lexical knowledge base created using WordNet. More recent studies have shifted towards using machine learning techniques for WSD. These approaches make use of both annotated and unannotated datasets, with SemCor being the most commonly used one. It is a dataset containing over 220,000 words that were manually annotated with senses from WordNet (Miller et al., 1993). One widely used method involves comparing contextual word embeddings with sense embeddings. This method requires either a corpus annotated with word senses (such as SemCor) or a word net (e.g., Loureiro and Jorge 2019; Levine et al. 2020).

Visual Word Sense Disambiguation (VWSD) can be considered as a multimodal extension of the conventional Word Sense Disambiguation task. Being a relatively new topic, there are not many available studies on this subject. An initial version of the task have been introduced by Barnard and Johnson (2005), where the image was used as an additional resource to disambiguate the sense of the word. Researchers employ statistical models to extract features from the co-occurrence of image regions and nouns in context. Most VWSD tasks involves identifying the sense of a word depicted in an image, given the image and the word. Recent studies on VWSD use a supervised WSD approach. For instance, Gella et al. (2016) generated two sense representations for each candidate sense, one based on text and the other on image. The researchers then applied an adapted Lesk Algorithm to measure the similarity between the image and senses in the sense inventory to determine the correct sense of the target word.

The SemEval 2023 VWSD task is unique in the sense that: given a word in its context and a list of images to choose from, it requires participants to select the correct image. Other WSD tasks, instead, typically ask to identify the textual sense of a word depicted in an image.

3 System Overview

The participants of the shared task were provided with two inputs: a text file and a folder that contains all candidate images. At each line, the text file comprises: a) the target word to disambiguate (e.g., *Andromeda* in Figure 1); b) the target word in context (e.g., *Andromeda tree* in Figure 1); c) the 10 candidate image references for that word.

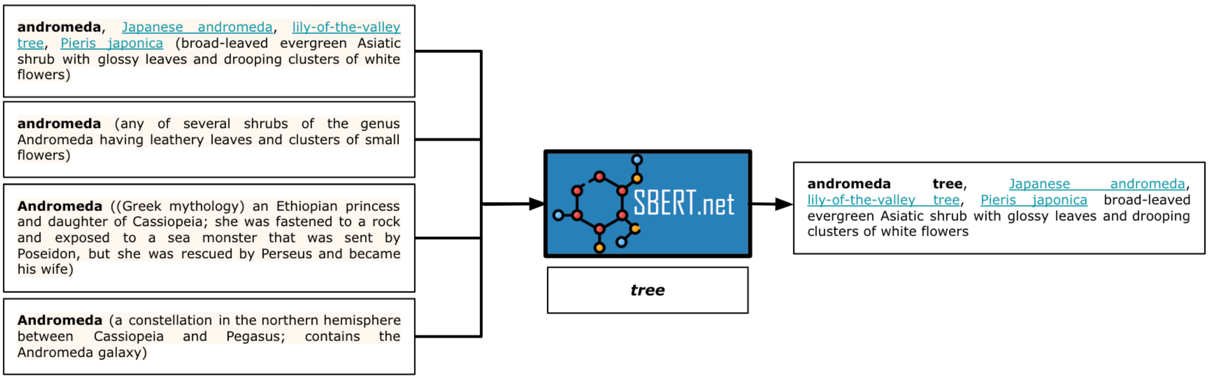
The linguistic context in this task is typically short, comprising one or two words. Moreover, around 18% of instances in the train set involve the following context words: *family*, *genus*, *species*, *phylum*, *class*, *order*. For some of the instances, the challenge is not textual ambiguity but selecting the correct image based on the limited textual information provided. For example, “malaxis genus” refers to a specific genus of orchid with only one sense, making disambiguation unnecessary. For these cases, the challenge lays in selecting the correct image from ten candidate options rather than disambiguation.

Due to the limitation of the linguistic context for disambiguation, we propose a two-step strategy that strongly rely on text augmentation. Firstly (3.1), we augmented the textual context using the external resources of WordNet and Wikipedia, leveraging sentence encoders to select the correct sense. Secondly (3.2), we make use of language-vision models to identify the correct image based on the extended context. Text augmentation has been found to be advantageous for the task.

3.1 Text Augmentation

To integrate the limited textual context, we leverage two external resources. Given the high degree of ambiguity of target words, the first resource employed is WordNet (Miller, 1994). WordNet aggregates word senses in synsets, with the related words and gloss. Figure 2 shows the different synsets (in bold) for the word *Andromeda*: each synset comprises the part-of-speech (in red), the related words (in blue) and the gloss (among brackets).

The idea here is to take the correct synsets



andromeda, Japanese andromeda, lily-of-the-valley tree, Pieris japonica (broad-leaved evergreen Asiatic shrub with glossy leaves and drooping clusters of white flowers)

andromeda (any of several shrubs of the genus Andromeda having leathery leaves and clusters of small flowers)

Andromeda ((Greek mythology) an Ethiopian princess and daughter of Cassiopeia; she was fastened to a rock and exposed to a sea monster that was sent by Poseidon, but she was rescued by Perseus and became his wife)

Andromeda (a constellation in the northern hemisphere between Cassiopeia and Pegasus; contains the Andromeda galaxy)

andromeda tree, Japanese andromeda, lily-of-the-valley tree, Pieris japonica broad-leaved evergreen Asiatic shrub with glossy leaves and drooping clusters of white flowers

Figure 2: Example of augmentation with WordNet: the target word *Andromeda* has 4 synsets and glosses, the context is *tree*.

(related words and gloss) as textual augmentation. However, almost every target word has different possible synsets. To choose the correct synsets for augmentation, we use the small context provided and Sentence Transformers (Reimers and Gurevych, 2019). Synsets are encoded using sBERT and compared to the given context (e.g., *tree* in *Andromeda tree*). Cosine similarity is then computed between all the synset embeddings and the context word embedding: the synset with the highest similarity to the context is selected for augmentation.

Alongside WordNet, Wikipedia has been employed as a source for augmentation. Specifically, the Wikipedia page summary of the keyword has been used for text augmentations. A page summary of Wikipedia contains an brief introduction to the subject and a summary of the page's most important contents. Wikipedia as a source of augmentation has been employed in two scenarios: when no synsets are found in WordNet and when the ambiguity is easy to resolve. More specifically, ambiguity was considered easy to solve when the phrase refers to the taxonomy name of an animal or a plant. These phrases have context word such as *species*, *genus*, *herb*). For these instances, we believe the encyclopedic knowledge from Wikipedia would have been more beneficial and straightforward than the WordNet strategy.

For the Italian set of the test, we retrieve target word definitions from an Italian dictionary[2]. If the word is not listed in the dictionary, its definition is sourced from Wikipedia as for English.

3.2 Image Retrieval

Our system is rather simple and straightforward. Given a keyword w, a context c, and a set of images I, we use a language-vision model to determine which image represents the sense of keyword is used in the context:

$$i_t = \arg\max(t \cdot (i^1, i^2, \ldots i^{10})) \tag{1}$$

where i_t represents the target image, t represents the text embeddings and i_n represents the image embeddings.

For the English dataset, we experimented with ViLT (Vision-and-language Transformer, Kim et al. 2021) and CLIP (Contrastive Language-Image Pre-Training, Radford et al. 2021) to find the correct image using keyword and enriched context. Based on transformer architecture, ViLT is a single-stream language-vision model, pre-trained on tasks of different modalities to effectively learn the relationships between images and text. CLIP is a language vision model trained on a contrastive learning task, where given a batch of N image-text pairs, CLIP is trained to predict which of the N^2 possible image-text pairings are actually matched. Unlike the single-stream architecture of ViLT, CLIP features separate encoders for each modality, with shallow interaction between them in the end. For the Italian dataset, we adopted Italian CLIP (Bianchi et al., 2021), which is an Italian version of CLIP model built upon Italian BERT and Vision transformer (Dosovitskiy et al., 2020).

4 Experimental Setup

For text augmentation, our approach combines WordNet and Sentence Transformers (sBERT). To access WordNet, NLTK toolkit has been employed.

[2] https://www.dizionario-italiano.it/

For sBERT, we conducted a preliminary research to identify the best pre-trained sentence encoder for our augmentation strategy. For the 16 samples in the trial set, we manually annotated the correct synsets in WordNet, then testing different sentence encoders. The best performing one, all-mpnet-base-v2, achieves an accuracy of 0.625 (10/16 samples). We consider this result satisfying for augmentation, either considering the extremely short context and the large number of synsets for common words (e.g., *bank*).

For the multimodal match, we conducted experiments in zero-shot settings. In the zero-shot setting, we ranked the images using the vanilla CLIP model based on clip-vit-base-patch32 and the ViLT model based on vilt-b32-finetuned-coco, following the approach described in Section 3. The maximum text length was set to 40 for ViLT and 77 for CLIP. The image resolution was 384 × 640 with Patch projection of 12 × 20 = 240 patches for the ViLT model and 1440 × 1810 for the CLIP model.

For each dataset, we use the augmented/unaugmented text with 10 images and ranked them according to their similarity (image, text) pair according to CLIP or ViLT. In detail, we first compute the word embedding of the text and the feature embedding of the 10 images with respect to text. Then, we ranked these 10 images based on the similarity scores of these embeddings.

5 Results

Two evaluation metrics are used to measure the performance of our model: hit rate at 1 and mean reciprocal rank. The *hit rate at 1* evaluates the accuracy of the first ranked image, with a score of 1 if the correct image is in the first position and 0 otherwise. The *mean reciprocal rank* is the average of the reciprocals of the rank position of the correct image (e.g., if the correct image is ranked second, the reciprocal rank would be 0.5). The baseline model provided by the task organizers is very similar to our system. They used CLIP to compute the text ("This is" + phrase to disambiguate) and image embeddings, and rank the candidate images based on the cosine similarity between the text and image embeddings.

The results shown in Table 1 indicate that the CLIP model outperforms the ViLT model. The best performance was achieved by the zero-shot CLIP model with augmentation (WordNet and Wikipedia), with a hit rate of 68.47 and a mean reciprocal rank of 79.17. Without text augmentation, the performance decreased to 58.96 for hit rate and 72.73 for mean reciprocal rank. For CLIP, we further try text augmentation only using Wikipedia: the low results (56.81 hit rate, 71.86 mrr) clearly demonstrate the positive effect of using WordNet on the task.

Model	hit rate	mrr
Baseline	61.32	74.65
ViLT	34.13	53.25
ViLT_aug	32.83	52.26
CLIP	58.96	72.73
CLIP_wiki_aug	56.81	71.86
CLIP_aug	**68.47**	**79.17**

Table 1: Results for CLIP and ViLT on the English dataset, with ("aug") and without augmentation. "wiki_aug"is CLIP using only Wikipedia as source for augmentation.

Model	hit rate	mrr
Baseline	23.61	44.59
CLIP	**37.70**	**55.66**
CLIP_aug	32.46	51.23

Table 2: Results for CLIP on the Italian dataset, with ("aug") and without augmentation, in terms of Hit Rate at 1 and Mean Reciprocal Rank (mrr).

While text augmentation produced an improvement in CLIP's overall performance, it actually worsened the performance of the ViLT model. This suggests that the CLIP is probably better equipped to handle diverse and potentially noisy text inputs while ViLT is more sensitive to changes in the input text and not able to handle text variations effectively. Also, CLIP uses large and seperate transformer embeddings for each modality which leads to a remarkable performance in text-to-image retrieval in zero-shot setting. In contrast, the single-stream architecture might be a reason of relatively poor performance in ViLT zero-shot setting.

Table 2 displays the results for the Italian dataset. Surprisingly, in the case of the Italian dataset, text augmentation did not improve the model's performance. In fact, both the hit rate and mrr decreased with the augmentation. This outcome, although contradictory to the results obtained for the English dataset, can be attributed to differences in the augmentation resources used. For the English dataset, the primary augmentation resources were

Wikipedia and WordNet, whereas for the Italian dataset, we utilized material from an online Italian dictionary. These results demonstrate that dictionary definitions are unsuitable for this type of task.

Our study highlights that text augmentation can benefit image retrieval, particularly for the CLIP model. In both languages, our system outperform the baseline system. However, it is crucial to evaluate the impact of text augmentation on each model, as it may not always lead to improved performance. We assessed the performance of both models using a trial dataset consisting of 16 instances, and our approach resulted in 10 accurate augmentations. The CLIP model correctly predicted 10 images, 9 of which were augmented with correct augmentation and 1 with incorrect augmentation. Interestingly, sometimes CLIP can retrieve the correct image using incorrect text augmentation. For example, "breaking wheel" refers to a medieval instrument of torture and execution, the text augmentation obtained from WordNet is "wheel around wheel somebody or something", which is obviously incorrect. However, CLIP model still selected the correct image. In contrast, the ViLT model only made 4 correct predictions, all of which were augmented with correct augmentation. We speculate the different performance of these two models may be due to the semantic encoder used in CLIP, which could enable the model to better capture the semantic relationship between words and images.

6 Conclusion

Overall, This project provides a brief overview of the word sense disambiguation (WSD) task and its visual counterpart (VWSD). It thoroughly discusses the differences and relationships between the two tasks. The VWSD task is formulated into two steps: text augmentation and image retrieval. The first step utilizes external resources such as WordNet. For the second step, the ViLT and CLIP models are experimented with, both achieving good results. We also tried to fine-tune the CLIP model, however fine-tuning did not help with model's performance.

7 Acknowledgement

We would like to express our sincere gratitude to Erasmus+ and Erasmus Mundus European Masters Program in Language and Communication Technologies (LCT) for providing support to students who participated in this project. We also thank Professor Raffaela Bernardi for her guidance throughout the workshop, Alberto Testoni and the students of course Grounded Language Processing for the feedbacks.

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 33–41, Athens, Greece. Association for Computational Linguistics.

Kobus Barnard and Matthew Johnson. 2005. Word sense disambiguation with pictures. *Artificial Intelligence*, 167(1-2):13–30.

Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Sri Lakshmi. 2021. Contrastive language-image pretraining for the italian language. *arXiv preprint arXiv:2108.08688*.

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. In *International Conference on Learning Representations*.

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.

Elisabetta Ježek. 2015. *The Lexicon: An Introduction*. Oxford University Press UK.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, Online. Association for Computational Linguistics.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy. Association for Computational Linguistics.

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*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 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.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.