POLITECNICO DI TORINO
Repository ISTITUZIONALE

PoliTo at SemEval-2023 Task 1: CLIP-based Visual-Word Sense Disambiguation Based on Back-Translation

Original
PoliTo at SemEval-2023 Task 1: CLIP-based Visual-Word Sense Disambiguation Based on Back-Translation / Vaiani, Lorenzo; Cagliero, Luca; Garza, Paolo. - ELETTRONICO. - (2023), pp. 1447-1453. (Intervento presentato al convegno SemEval-2023 (Workshop of ACL) tenutosi a Toronto (CAN) nel July 9–14, 2023) [10.18653/v1/2023.semeval-1.199].

Availability:
This version is available at: 11583/2982327 since: 2023-09-20T08:27:47Z

Publisher:
ACL Association for Computational Linguistics

Published
DOI:10.18653/v1/2023.semeval-1.199

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

09 May 2024

PoliTo at SemEval-2023 Task 1: CLIP-based Visual-Word Sense Disambiguation Based on Back-Translation

Lorenzo Vaiani and **Luca Cagliero** and **Paolo Garza**
Politecnico di Torino
{*lorenzo.vaiani,luca.cagliero,paolo.garza*}*@polito.it*

Abstract

Visual-Word Sense Disambiguation (V-WSD) entails resolving the linguistic ambiguity in a text by selecting a clarifying image from a set of (potentially misleading) candidates. In this paper, we address V-WSD using a state-of-the-art Image-Text Retrieval system, namely CLIP. We propose to alleviate the linguistic ambiguity across multiple domains and languages via text and image augmentation. To augment the textual content we rely on back-translation with the aid of a variety of auxiliary languages. The approach based on fine-tuning CLIP on the full phrases is effective in accurately disambiguating words and incorporating back-translation enhances the system's robustness and performance on the test samples written in Indo-European languages.

1 Introduction

Visual-Word Sense Disambiguation (V-WSD) is a seminal task in the field of multimodal learning involving both natural language processing and computer vision (Raganato et al., 2023). It entails the identification of the correct sense of a given text containing ambiguous words by selecting a clarifying image among a set of candidates, possibly including misleading options. Tackling the aforesaid task is relevant to address well-known real-world challenges such as multimodal machine translation (Barrault et al., 2018; Gella et al., 2019), image caption generation (Wang et al., 2020a; Pramanick et al., 2022), visual search and object recognition (Yao et al., 2019; Zou et al., 2023).

The key issues in V-WSD are the linguistic ambiguity of the expressions occurring in the text, which may subtend multiple senses, and the high variability of the candidate images across different domains, which makes the problem of image retrieval non-trivial. To address the above issues, we propose a novel approach to V-WSD based on CLIP (Radford et al., 2021), i.e., a state-of-the-art vision-language model trained on massive data consisting of image-caption pairs. Leveraging the outstanding capability of CLIP to capture associations between visual and textual content, we first fine-tune the pretrained model on the raw V-WSD data to assess model robustness across multiple domains and languages. Next, we explore the use of *data augmentation* on both text and images, where text is augmented via *back-translation*. Considering text translations in multiple languages has the twofold aim of enriching the model with new text variants generated via machine translation and making the CLIP model suitable for multilingual scenarios. We also empirically analyze the effect of considering different snippets of the input text, i.e., the ambiguous words only, the main topic description, or the full phrase.

The results confirm the suitability of the fine-tuned CLIP version to effectively tackle the V-WSD by extracting the key information from the full phrases. Specifically, they show a performance boost on the family of Indo-European languages that are predominant in the training data whereas confirming the expected issues with underrepresented Indo-Iranian languages.

Our results achieve the 40th rank out of 91 submissions for the English test split, the 28th rank out of 58 submissions for the Farsi test split, and the 29th rank out of 64 submissions for the Italian test split.

The rest of the paper is organized as follows. Section 2 formalizes the addressed task. Section 3 reviews the related literature. Section 4 presents the proposed methodology. Section 5 describes the preliminary results, whereas Section 6 draws the conclusions of the work.

2 Task Overview

The V-WSD task frames an image-text retrieval problem, where the goal is to retrieve a target image given a textual query. The challenge is based on a bespoke dataset that comprises 12869 instances,

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1447–1453
July 13-14, 2023 ©2023 Association for Computational Linguistics

Word	Count (overall)	Count (as target word)
Frequent words		
genus	1238	0
family	385	0
herb	125	1
order	114	0
shrub	98	0
Frequent ambiguous words		
drive	7	5
catch	5	4
roll	5	4
electronics	3	3
groom	5	3

Table 1: Examples of frequent and ambiguous words occurring in the training dataset.

each one consisting of a brief text accompanied by a set of 10 associated images. Each piece of text contains an ambiguous word, which is known to challenge's participants, and the images are selected from a set of 12999 pictures. Notice that for each instance the images in the pool can range from being completely extraneous to the textual content to pertinent to a specific aspect of it, such as the ambiguous words, and/or the remaining text portion.

Table 1 reports some examples of words occurring in the training data, differentiating between *frequent* words and *ambiguous* ones. Frequent words such as "*genus*", "*family*" and "*herb*" are those characterized by a relatively high-frequency count and commonly refer to general topics, e.g., the fauna and flora types, and are unlikely to manifest as ambiguous words. Conversely, most frequent ambiguous terms such as "*drive*" are commonly not topic-specific, rarely appear in the training data, and are, in most of the instances they occur, the target words yielding a kind of ambiguity.

The V-WSD task also takes language diversity into consideration. Specifically, although the training corpus is exclusively written in the English language, the test set includes text snippets written in English, Farsi, and Italian. Hence, effectively handling multilingual textual data is particularly relevant.

3 Related Works

The V-WSD task concerns the polysemy between textual and visual information thus requiring a joint effort of computer vision and natural language processing experts. A word can be deemed as ambiguous if expresses multiple concepts and thus can be associated with different images. The ultimate goal of the V-WSD task is to provide a comprehensive and coherent understanding of multimedia data, reconciling the ambiguity between textual and visual content.

We can disentangle the V-WSD task into two well-known sub-problems, which need to be jointly addressed.

3.1 Word Sense Disambiguation

Word sense disambiguation (WSD) is a subfield of computational linguistics that focuses on determining the correct meaning of a word in a given context (Bevilacqua et al., 2021). WSD can be addressed by means of different approaches among which knowledge-based (Wang et al., 2020b), supervised methods (Maru et al., 2019) and hybrid methods (Bevilacqua and Navigli, 2020). In our proposed approach, we leverage supervised deep learning methods to capture the semantic and contextual word relationships. Such information is then exploited to address word disambiguation.

Textual data augmentation is established for improving the WSD performance (Lin and Giambi, 2021). Specifically, back-translation (Edunov et al., 2018) stands out as a noteworthy augmentation strategy. It entails first translating a piece of text into a different language and then translating it back into the original language. Back-translation is instrumental for word disambiguation because

- Likely introduces new word/phrase variations (e.g., synonyms, antonyms, and related words) that enrich the model with additional contextual information.
- Improves the robustness/accuracy of WSD algorithms by exposing them to more diversified training instances, e.g., different linguistic structures and cultural variations.

WSD has primary importance in V-WSD because a correct interpretation of the word context of usage is crucial for correctly retrieving the related images.

3.2 Image-Text Retrieval

Image-Text Retrieval is a task that involves both vision and language and consists in retrieving relevant samples from one modality based on queries

expressed in the other modality (Cao et al., 2022). V-WSD task is an example of text-to-image (T2I) retrieval where text queries are used to select the corresponding image.

A viable solution to this challenge is the use of cross-modal embedding models, whose aim is to learn a common representation space for both images and text. Embedding learning can be accomplished via either siamese networks (Fang et al., 2015; Si et al., 2018) or by using models trained to project both modalities into a joint embedding space (Yuan et al., 2021). Among them, the Contrastive Language-Image Pre-Training (CLIP) model (Radford et al., 2021) has been shown to be particularly effective in mapping images and textual descriptions into a common embedding space. The text-image similarity is simply measured via the dot product of the corresponding embeddings. In the present work, we leverage a large-scale pretrained CLIP model for tackling the T2I as achieved state-of-the-art performance in multiple scenarios.

4 Methodology

In this section, we describe the core elements of the methodology proposed to tackle the V-WSD task. Hereafter, we will separately discuss the definition of the *text snippet* to provide as an input to the CLIP model (see the definition in Section 4.1 and the related experiments in Section 5.1) and the *data augmentation* procedure (see the details in Section 4.2 and the related experiments in Section 5.2).

4.1 Text snippet definition

We explore the use of the following text snippets:

- *Full Phrase* (FP), which includes both the ambiguous word and its surrounding context;
- *Ambiguous Word* (AW), which comprises the target word only (disregarding the context);
- *Main Topic* (MT), which consists of the original piece of text excluding the ambiguous word.

We use the pre-trained version of the CLIP model to process the different text input versions and quantify their relevance to the V-WSD task. Specifically, for each of the above-mentioned text snippets we first compute the similarity between its encoding and the embedding of each candidate image. Then, we select the picture with the highest similarity score as the target image. The key idea is to gain a better understanding of which textual features mainly contribute to WSD.

To further improve the performance of the CLIP model on the V-WSD task, we investigated the potential benefits of fine-tuning the model on a subset of the provided dataset that is particularly rich in ambiguities. As indicated by the statistical analysis presented in Table 1, several high-frequency words in the dataset suggest the presence of recurring topics. Therefore, a model that is fine-tuned on such an ambiguous dataset would be better suited for disambiguating the target word, as it would be trained to capture the nuances of the specific language used within the dataset. By fine-tuning the model on ambiguous terms, we could potentially improve its ability to discriminate between different senses of the same word and produce more accurate predictions.

4.2 Data Augmentation

Recent advancements in deep learning have demonstrated the effectiveness of data augmentation in improving model performance in a variety of tasks (Shorten and Khoshgoftaar, 2019; Bayer et al., 2022). We explore the use of data augmentation to enhance the quality and robustness of the trained V-WSD models. Specifically, we apply various transformations to both the input text and the associated images, generating new data instances that likely capture complementary meanings and contexts of use of the target words.

Text augmentation We employ back-translation to create new phrases similar to the original pieces of text. Specifically, we used Farsi, Italian, French, and German languages to translate the original text and then back-translate it to English. Italian and Farsi languages have been selected because included in test samples, while German and French languages have been selected to enhance language diversity since they belong to the Germanic and Romance language subgroups respectively.

To augment textual data, we use pretrained machine translation models from English to intermediate languages, and vice versa, which are provided by Helsinki University[1] for Italian, German, and French, and by PerisanNLP[2] for Farsi. Specifically, back-translation is applied to 40% of the training samples, with each intermediate language selected

[1] https://huggingface.co/Helsinki-NLP
[2] https://huggingface.co/persiannlp

Training Prompt	Inference Prompt	Hit Rate	MRR
-	**FP**	79.80%	86.96%
-	**MT**	50.01%	65.11%
-	**AW**	70.86%	80.33%
FP	**FP**	**83.84%**	**89.77%**
FP	**MT**	57.58%	71.03%
FP	**AW**	75.64%	83.73%
MT	**FP**	82.67%	89.04%
MT	**MT**	63.09%	75.37%
MT	**AW**	74.63%	83.09%
AW	**FP**	83.14%	89.21%
AW	**MT**	56.64%	70.59%
AW	**AW**	76.34%	84.26%

Table 2: Results obtained varying the textual prompt.

Textual Augmen.	Visual Augmen.	Hit Rate	MRR
-	-	**83.84%**	**89.77%**
-	✓	83.37%	89.52%
✓	-	83.57%	89.62%
✓	✓	83.41%	89.59%

Table 3: Results obtained exploiting textual/visual augmentation techniques.

with equal probability. Considering the ambiguity of the texts, we decided to use most of the texts in their original form to avoid introducing excessive errors due to the use of translation models. This back-translation pipeline is applied at each epoch, thus the textual part of the same training sample can appear in different versions during the whole training step.

Visual augmentation We generate new images from the original picture representing the ambiguous sentence from different perspectives. We apply the following transformations to the training samples: horizontal flipping, color jittering, and grey scaling. Each transformation is applied with a probability of 20%, regardless of the selection of the other available transformations. The aim is to improve the robustness of the visual encoder by enhancing its ability to handle different variations of the input text and images.

5 Experimental Results

We empirically analyze the impact of each core element of the proposed methodology on the V-WSD dataset. To this end, we use the official pretrained large version of the CLIP model[3] throughout the experimental session. To ensure the robustness and generalizability of the model, we randomly select a validation set comprising 20% of the labeled training data to monitor the model's performance during training. We subsequently evaluate the test set by selecting the best performing model according to the results obtained on the validation set.

Evaluation metrics We evaluate our models using the official V-WSD metrics, which are also established for Image-Text retrieval, i.e., hit rate and mean reciprocal rank (MRR). Hit rate measures the proportion of correct predictions made by the model, whereas MRR provides a measure of the ranking of the correct prediction. We test our trained models on the provided test splits using the official evaluation metrics as in the training phase. In particular, we evaluate the English split as is, while for the Farsi and Italian splits, we translate the input text into English using the same publicly available transformer-based models that we use for back-translation during training.

Algorithms' configurations. We train our models for 30 epochs to minimize a symmetric cross-entropy loss with Adam (Kingma and Ba, 2015), using a batch size of 16 and 10^{-7} as learning rate. For the sake of reproducibility, we make the code publicly available[4].

Hardware Settings. We run the experiments on a machine equipped with AMD® Ryzen 9® 3950X CPU, Nvidia® RTX 3090 GPU, and 128 GB of RAM running Ubuntu 21.10.

5.1 Effect of the input text snippet

Table 2 reports the V-WSD model's performance (achieved on the validation set) leveraging different textual inputs, i.e., Full Phrase (FP), Ambiguous Word only (AW), and Main Topic only (MT). The pretrained model achieves the best results while taking FP as input. Conversely, MT yields the worst performance. The results achieved using AW are slightly worse than those of FP. The aforesaid result holds true in the fine-tuned versions of the model, where using FP as input during the inference phase consistently leads to the best overall results.

[3] https://huggingface.co/openai/clip-vit-large-patch14

[4] The GitHub project repository is available at https://github.com/VaianiLorenzo/VWSD

	EN		FA		IT		AVG	
	Hit Rate	**MRR**	**Hit Rate**	**MRR**	**Hit Rate**	**MRR**	**Hit Rate**	**MRR**
Baseline	60.47%	73.88%	28.50%	46.70%	22.62%	42.61%	37.20%	54.39%
Submission #1	**65.23%**	77.81%	**41.00%**	**56.64%**	55.41%	69.94%	**53.88%**	**68.13%**
Submission #2	**65.23%**	**78.30%**	38.50%	55.32%	**56.72%**	**70.77%**	53.48%	**68.13%**

Table 4: Results summary (test set).

5.2 Effect of the data augmentation

Table 3 presents the validation set results of the V-WSD model while different data augmentation techniques are incorporated. Surprisingly, we do not observe any improvement in performance, regardless of the type of data augmentation employed. One possible explanation for the observed behavior could be that the data augmentation techniques employed do not offer additional meaningful information to the V-WSD model applied on an English-language validation split. The augmented data may have led to an increase in noise or redundancy, which could have resulted in slight performance deterioration. However, the exact underlying causes of this behavior require further investigation and analysis.

Table 4 reports the results achieved by the two official submissions. We fine-tune the CLIP model using the full phrase as input text snippet on the entire training dataset without data augmentation (namely, *Submission* #1) and using both textual and visual augmentation techniques (*Submission* #2). To allow a direct language-specific comparison, we also report the results obtained by the task organizers' baseline, which is also CLIP-based, as well as the result of our submissions on all test splits. Overall, our submission results outperform the proposed baseline. This performance gap confirms once again the usefulness of fine-tuning CLIP on ambiguous text snippets. Moreover, focusing on the English test split, it is worth noticing that the result are worse than those obtained on the validation set. This could be due to the discrepancy in recurrent topics between train and test splits, which could have made the model less effective in handling certain types of ambiguous sentences.

The results achieved on the official test instances reverse the situation compared to the preliminary outcomes on the validation set: data augmentation slightly improves the MRR performance on the English test set and the hit rate and MRR scores on the Italian test instances. Conversely, on the Farsi test instances data augmentation causes a decrease in performance, especially in the hit rate. This is likely caused by the strong dependence holding between the back-translation effectiveness and both the quality of the employed machine translation model and the similarity between the source and target languages. In fact, Italian, French, and German share many syntactic and linguistic features with English. Oppositely, Farsi belongs to the Iranian branch of the Indo-Iranian language family, which is substantially different from Indo-European languages.

6 Conclusions and Discussion

The proposed visual-word sense disambiguation (V-WSD) system has demonstrated a remarkable ability to accurately disambiguate lexical items and select the corresponding image. Our empirical results have highlighted the robustness of the system when pieces of text written in Indo-European languages are translated to English and given as input. The preliminary results underscore the potential of leveraging the CLIP model for natural language processing tasks that demand multimodal understanding.

Future research endeavors will entail exploring the use of a multilingual CLIP model to obviate the translation process during inference and broaden the system's scope to accommodate additional language families, including more Indo-Iranian and other underrepresented languages. We also plan to observe the effect of back-translation individually for each intermediate language, also varying its aggressiveness. Ultimately, our proposed system has far-reaching applications, including multimodal machine translation, image caption generation, and visual question answering.

Acknowledgements

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPO-

NENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

This study was carried out within the MICS (Made in Italy – Circular and Sustainable) Extended Partnership and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1551.11-10-2022, PE00000004). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

This work has been also partially supported by the SmartData@PoliTO center on Big Data and Data Science.

References

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A survey on data augmentation for text classification. *ACM Comput. Surv.*, 55(7).

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*. International Joint Conference on Artificial Intelligence, Inc.

Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. 2022. Image-text retrieval: A survey on recent research and development. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5410–5417. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482.

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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Guan-Ting Lin and Manuel Giambi. 2021. Context-gloss augmentation for improving word sense disambiguation. *arXiv preprint arXiv:2110.07174*.

Marco Maru, Federico Scozzafava, Federico Martelli, and Roberto Navigli. 2019. SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations. 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 3534–3540, Hong Kong, China. Association for Computational Linguistics.

Pradip Pramanick, Chayan Sarkar, Sayan Paul, Rudra dev Roychoudhury, and Brojeshwar Bhowmick. 2022. Doro: Disambiguation of referred object for embodied agents. *IEEE Robotics and Automation Letters*, 7(4):10826–10833.

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.

Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. 2018. Dual attention matching network for context-aware feature sequence based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5363–5372.

Haoran Wang, Yue Zhang, Xiaosheng Yu, et al. 2020a. An overview of image caption generation methods. *Computational intelligence and neuroscience*, 2020.

Yinglin Wang, Ming Wang, and Hamido Fujita. 2020b. Word sense disambiguation: A comprehensive knowledge exploitation framework. *Knowledge-Based Systems*, 190:105030.

Yazhou Yao, Zeren Sun, Fumin Shen, Li Liu, Limin Wang, Fan Zhu, Lizhong Ding, Gangshan Wu, and Ling Shao. 2019. Dynamically visual disambiguation of keyword-based image search. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 996–1002. International Joint Conferences on Artificial Intelligence Organization.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. 2021. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6995–7004.

Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. 2023. Object detection in 20 years: A survey. *Proceedings of the IEEE*, pages 1–20.