TAM of SCNU at SemEval-2023 Task 1: FCLL: A Fine-grained Contrastive Language-Image Learning Model for Cross-language Visual Word Sense Disambiguation

Qihao Yang†, Yong Li†, Xuelin Wang‡, Shunhao Li†, Tianyong Hao†*
†School of Computer Science, South China Normal University, Guangzhou, China
‡College of Chinese Language and Culture, Jinan University, Guangzhou, China
{charlesyeung, lycutter, haoty}@m.scnu.edu.cn
wangxuelin@stu2022.jnu.edu.cn, lishunhao99@foxmail.com

Abstract

Visual Word Sense Disambiguation (WSD), as a fine-grained image-text retrieval task, aims to identify the images that are relevant to ambiguous target words or phrases. However, the difficulties of limited contextual information and cross-linguistic background knowledge in text processing make this task challenging. To alleviate this issue, we propose a **F**ine-grained **C**ontrastive **L**anguage-Image **L**earning (FCLL) model, which learns fine-grained image-text knowledge by employing a new fine-grained contrastive learning mechanism and enriches contextual information by establishing relationship between concepts and sentences. In addition, a new multimodal-multilingual knowledge base involving ambiguous target words is constructed for visual WSD. Experiment results on the benchmark datasets from SemEval-2023 Task 1 show that our FCLL ranks at the first in overall evaluation with an average H@1 of 72.56% and an average MRR of 82.22%. The results demonstrate that FCLL is effective in inference on fine-grained language-vision knowledge. Source codes and the knowledge base are publicly available at https://github.com/CharlesYang030/FCLL.

1 Introduction

The problem of word polysemy, first recognized in machine translation, is one of challenging tasks in Natural Language Processing (NLP). In the field of NLP, resolving word polysemy is generally regarded as Word Sense Disambiguation (WSD), and it still remains as one of the most challenging and pervasive linguistic phenomena in NLP at present (Bevilacqua et al., 2021). To promote the research of the issues, international competitions have been held such as SemEval[1] to advance the state-of-the-art in word sense analysis and to help create high-quality annotated datasets in WSD. Since SemEval-2023, visual WSD has been introduced and it contains three tracks including English, Farsi and Italian (Raganato et al., 2023). Given a target word and some limited textual context, the visual WSD task is to select the one which relevant to the intended meaning of the word among a set of candidate images.

Theoretically, the general WSD task aims at making explicit the semantics of a word in context by identifying the most suitable meaning (called sense) from a pre-defined sense inventory (Bevilacqua et al., 2021), which only involves words and senses. Differently, visual WSD requires models to learn more fine-grained image-text knowledge to establish the relationship among words, senses and images. Meanwhile, the limitations of visual WSD are significant. Because it is difficult in most cases to determine the specific sense of a word from a phrase containing few words without additional contextual information, and even more difficult to further identify the images related to the word. In addition, words in text of different languages may have different senses, and the syntactic and structural differences among languages make models ineffective during textual processing. Therefore, cross-language background knowledge is another major challenge in visual WSD.

In this paper, we propose a **F**ine-grained **C**ontrastive **L**anguage-Image **L**earning (FCLL) model for learning fine-grained image-text knowledge. The main contributions of this work lies on three aspects:

1. A new model named FCLL is proposed, in which a subtlety discriminative text encoder and an image encoder are designed from text-inferred images and image-inferred texts perspective respectively.
2. A new visual WSD knowledge base named V-WSD KB as an easily expandable multimodal-multilingual knowledge base is constructed. It contains three languages, 12956 ambiguous

[*] Corresponding author.
[1] https://semeval.github.io/

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 506–511
July 13-14, 2023 ©2023 Association for Computational Linguistics

target words, 20904 senses and 97267 relevant images.

3. A new sense auto-complementing strategy is proposed, where an exclusive sense encoder is activated to complement English contextual information to phrases.

2 Related Work

In recent years, the dominant approaches in general WSD have included KB-based and supervised methods. Most KB-based models, such as SyntagRank (Scozzafava et al., 2020) and SREF$_{KB}$ (Wang and Wang, 2020) , employ specific rank algorithms and retrieval rules. Most supervised approaches (e.g., SensEmBERT (Scarlini et al., 2020) and BEM (Blevins and Zettlemoyer, 2020)) are based on neural network systems that use pre-trained models. Besides, several studies (e.g., (Gella et al., 2019)) have utilized visual features to solve the problem of textual WSD. However, these models do not involve fine-grained image-text knowledge and cannot be directly applied to visual WSD.

In vision-language pre-training, contrastive learning (Hadsell et al., 2006) focuses on learning common features between similar instances. CLIP (Radford et al., 2021) employs a classical two-stream structure using more than 400 million image-caption pairs collected from the Internet as training data, which is equipped with an outstanding performance for zero-shot. BLIP (Li et al., 2022) is an extension of CLIP, learning from noisy image-text pairs by bootstrapping the captions. Moreover, FLAVA (Singh et al., 2022) is an alignment model, which obtains image-text features separately using unimodal encoders, followed by multimodal pre-training. However, most of the existing multimodal pre-trained models use manually managed datasets that are coarse-grained, such as COCO (Lin et al., 2014) and Flickr-30 (Plummer et al., 2015). Furthermore, these models are studied in the monolingual background.

3 Method

3.1 The FCLL Model

We propose the FCLL model for learning multimodal-multilingual fine-grained image-text knowledge. Figure 1 illustrates the learning framework of the model, which is composed by four modules including a fine-grained text encoder, a fine-grained image encoder, a fine-grained image-text matching predictor and a word sense auto-

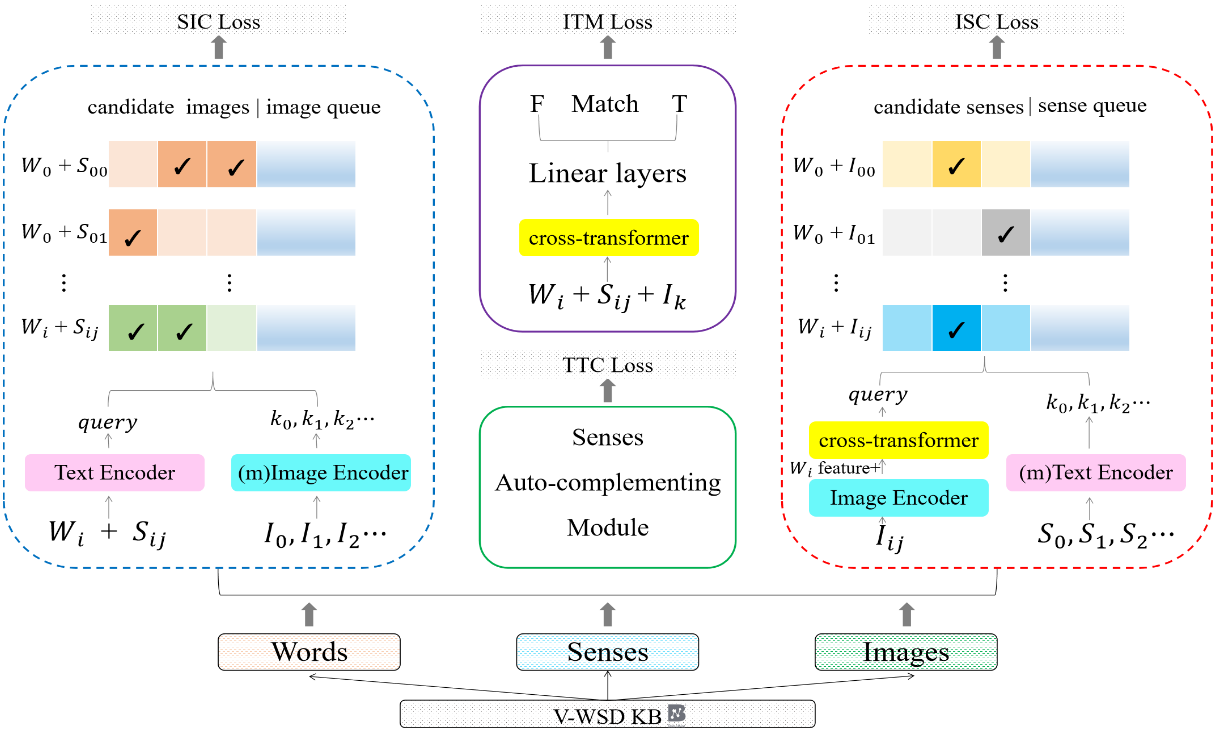

Figure 1: Learning framework of FCLL. W_i represents the i-th ambiguous target word in a mini-batch, S_{ij}/I_{ij} denotes the j-th sense or image corresponding to W_i, and the encoder starting with "(m)" illustrates the momentum encoder.

complementing module. Inspired by MoCo (He et al., 2020), the FCLL model maintains an image queue and a text queue separately.

Fine-grained Text Encoder (the blue dashed box). All the relevant images of ambiguous target words in a mini-batch are retrieved, and a text encoder based on CLIP promotes semantic consistency between the senses guided by these words and the images from the text-inferred image perspective. Then, these images are updated to an image queue with length 80000.

Fine-grained Image Encoder (the red dashed box). Differently, all the senses of ambiguous target words in a mini-batch are retrieved, and an image encoder based on CLIP activates semantic consistency among the images guided by these words and the senses from the image-inferred text perspective. Then, these senses are updated to a text queue with length 20000.

Fine-grained Image-Text Matching Predictor (the solid purple box). A transformer connected by a simple classifier is trained to fuse the multimodal features of the "words + senses + images" combinations. The matching predictor as well as facilitates the contrastive learning from the perspective of feature fusion.

3.2 The Word Sense Auto-complementing

Theoretically, the length of phrases affects the performance of a visual WSD model to some extent. Therefore, an exclusive sense encoder is trained for complementing a desirable sense to a phrase. As shown in Figure 2, FCLL removes the stop words of each original sense (called the full sense) and combines any one of the remaining words with their corresponding ambiguous target word to form a concept, and subsequently promotes semantic consistency between the concept and the full sense. Note that all the non-English texts in test set are translated to English by Google Translate, and the candidate synsets (synonym sets) are provided by BabelNet (Navigli and Ponzetto, 2010; Navigli et al., 2021).

3.3 Loss Function

Contrastive losses (Hadsell et al., 2006) is used to calculate the similarities of two vectors in a representation space. Based on ALBEF (Li et al., 2021), we define three contrastive losses including Sense$_{atw}$-Image Contrastive (SIC) loss, Image$_{atw}$-Sense Contrastive (ISC) loss and Text$_{concept}$-Text$_{full}$ Contrastive (TTC) loss. In the SIC loss, the combinations of the senses guided by an ambiguous target word (*atw*) and positive images are used as two vectors. In the ISC loss, the combinations of images guided by an ambiguous target word and positive senses are used as two vectors. Similarly, in the TTC loss, the combinations of concepts and full texts are used as two vectors. In addition, we follow BLIP and define a cross entropy loss named Image-Text Matching loss, which is a binary classification task essentially.

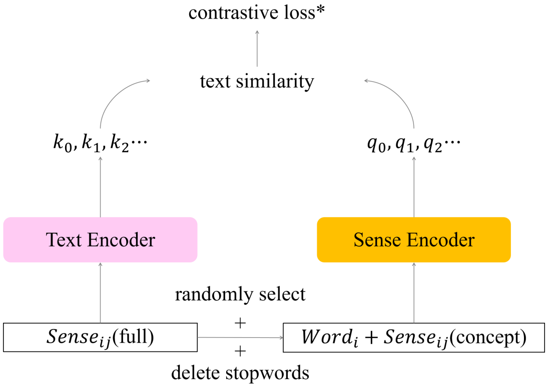

Figure 2: Sense Auto-complementing Module. $Word_i$ denotes the i-th ambiguous target word in a mini-batch and $Sense_{ij}$ represents the j-th sense regarding $Word_i$. "$*$" indicates the unit matrix is used as the label.

3.4 The Visual WSD Knowledge Base Construction

Empirically, the datasets provided by SemEval-2023 Task 1 suffer from low data volume, monolingual English, and the lack of contextual information, which is further discussed in Section 4.1. To this end, we firstly select ambiguous target words as queries based on officially released datasets and then crawl all the synsets of these queries online by the API guidelines[2] from BabelNet. Here we choose English, Farsi and Italian. The resource of synsets comes from WordNet (Miller et al., 1990), and the part-of-speech is set as noun. For each sense, several relevant images are displayed in BabelNet. We employ a top-5 strategy, i.e., only the top five images are collected (due to the format and legality of images, less than five images can be collected in some cases). After nearly two months of collection and validation, V-WSD KB contains 12,956 multilingual ambiguous target words, 20,904 English senses, and 97,267 relevant images. The maximum number of senses for an ambiguous target word is 23, the minimum number

[2] https://babelnet.org/guide

is 1, and the mean number is 1.613. The maximum number of images for a sense is 5, the minimum number is 1, and the mean number is 4.653. Note that SemEval-2023 Task 1 claims that participants are allowed to use external data sources. Accordingly, V-WSD KB is constructed based on BabelNet 5.2 under SapienzaNLP license agreement and is used as complementary training data for FCLL.

4 Experiments and Results

4.1 Datasets

We use the training/test sets released by SemEval-2023 Task 1, and the distribution of ambiguous target words for each language in the datasets is shown in Table 1. Although the training set involves multiple languages, their proportions are severely imbalanced. Instead, only three languages are involved in the test set. Furthermore, there is only one data record in Farsi and even none in Italian in training set, and there is only one reduplication of target words in training/test sets.

Training set			Test set		
type	number	proportion	type	number	proportion
English	12825	99.658%	English	463	47.830%
Chinese	8	0.062%	Farsi	200	20.661%
Thai	2	0.015%	Italian	305	31.508%
Hindi	1	0.007%	-	-	-
Yiddish	1	0.007%	-	-	-
Japanese	1	0.007%	-	-	-
Greek	1	0.007%	-	-	-
Korean	1	0.007%	-	-	-
Farsi	1	0.007%	-	-	-
icon	21	0.163%	-	-	-
number	6	0.046%	-	-	-
symbol	1	0.007%	-	-	-
total : 12869			total : 968		

Table 1: The distribution of the official datasets for each language.

Table 2 shows the differences among training set, test set and V-WSD KB. Note that the structure of "atw-text-image" (i.e., word-sense-image) correspondence as "1-n-n" means that each ambiguous target word may correspond to multiple meanings, and each meaning also corresponds to multiple associated images.

4.2 Settings

Visual WSD is recognized as a ranking problem and it adopts Hit Rate at 1 (H@1) and Mean Reciprocal Rank (MRR) to measure model performance, which can be calculated by Eq. (1) and (2):

$$H@1 = \frac{\text{Number of Hits @1}}{Q} \tag{1}$$

$$MRR = \frac{1}{Q} \sum_{i}^{|Q|} \frac{1}{\text{Rank}_i} \tag{2}$$

Q indicates the total number of recommendation lists, and Number of Hits @1 represents the number of items in the first position of each recommendation list that are consistent with ground-truth, while Rank_i denotes the column position of the item in the i-th recommendation list that is consistent with ground-truth.

Our model was implemented on Pytorch and a single Nvidia A100 GPU in about 35 hours. We set the mini-batch size to 2, the initial learning rate to 0.0001, and the number of training epochs to 10. Furthermore, we use AdamW as the optimizer, and the weight decay of AdanW is 0.05. After each epoch, the model performs a cosine learning rate decay. We follow CLIP and specify an image resolution of 224×224, a maximum text length of 77, and a text that starts with "A photo of".

4.3 Results

There are 98 submissions in this visual WSD campaign. FCLL wins the first place on the leaderboard for the final average score in multiple languages. We collect the results of the official baseline and

Items	Training set	Test set	V-WSD KB
Num. of ambiguous target words (atw)	12869	968	12956
language of atw	English	English, Farsi, Italian	English, Farsi, Italian
Num. of texts	12869	968	20904
form of text composition	phrases composed by 2 words	phrases composed by 2~3 words	sentences
language of text	English	English, Farsi, Italian	English
Num. of images	12999	8100	97267
atw-text correspondence	1-1	1-1	1-n
text-image correspondence	1-1	1-1	1-n
atw-text-image correspondence	1-1-10	1-1-10	1-n-n
size	16.8GB	10.4GB	114GB

Table 2: The differences among training set, test set and V-WSD KB.

Models	Parameters	English		Farsi		Italian		Average	
		H@1(%)	MRR(%)	H@1(%)	MRR(%)	H@1(%)	MRR(%)	H@1(%)	MRR(%)
Baseline†24	-	60.475	73.876	28.500	46.697	22.622	42.606	37.199	54.393
Samsung Research China - Beijing†2	-	**84.017**	**89.558**	59.000	70.513	72.459	82.080	71.825	80.717
OPI, Poland†3	-	77.969	85.879	**64.000**	**74.387**	69.508	79.145	70.492	79.804
CLIP*	151M	56.371	70.398	52.000	65.580	54.754	69.370	54.375	68.449
BLIP^*_{COCO}	447M	57.667	72.043	51.000	64.588	58.032	71.034	55.566	69.222
$\text{BLIP}^*_{Flickr-30}$	447M	60.259	73.685	50.000	64.225	57.377	70.751	55.878	69.554
FLAVA*	242M	17.278	36.876	15.500	33.934	15.081	35.397	15.953	35.402
FCLL†1	189M	80.129	87.417	60.500	73.190	**77.049**	**86.047**	**72.559**	**82.218**

Table 3: Evaluation on the benchmark test set. "†" indicates the officially published result, the number following it represents the final ranking in the official average score, and "$*$" denotes zero-shot for reproduction using available codes of the baselines on our experiment environment.

Models	Fine-grained text enc.	Fine-grained image enc.	Matching predictor	Senses Auto-com	English		Farsi		Italian		Average	
					H@1(%)	MRR(%)	H@1(%)	MRR(%)	H@1(%)	MRR(%)	H@1(%)	MRR(%)
$\text{FCLL}_{no\ t-enc.}$	✗	✓	✓	✓	39.740 -40.389	60.801 -26.616	33.000 -27.500	54.796 -18.394	42.622 -34.427	62.653 -23.394	38.454 -34.105	59.417 -22.801
$\text{FCLL}_{no\ i-enc.}$	✓	✗	✓	✓	75.809 -4.32	84.820 -2.597	57.499 -3.001	71.922 -1.268	74.098 -2.951	84.101 -1.946	69.136 -3.423	80.281 -1.937
$\text{FCLL}_{no\ m-pre.}$	✓	✓	✗	✓	75.161 -4.968	84.324 -3.093	56.499 -4.001	71.100 -2.090	74.098 -2.951	83.982 -2.065	68.586 -3.973	79.802 -2.416
FCLL_{noSA}	✓	✓	✓	✗	60.475 -19.654	74.935 -12.482	45.500 -15.000	61.742 -11.448	60.983 -16.066	74.189 -11.858	55.652 -16.907	70.289 -11.929

Table 4: Ablation study on the benchmark test set. The "✗" indicates the removed module.

the top-3 models, and then we use CLIP, BLIP and FLAVA for zero-shot on the test set that has been translated to English. The comparison results are shown in Table 3. FCLL outperforms the baseline models and shows the flexibility to language changes, whose balanced performance is benefited from its fine-grained core modules. In addition, BLIP is comparable to CLIP in terms of average score. The former has 447M parameters, while the later has 151M. In contrast, FCLL has only 38M parameters more than the original CLIP and further advance the performance on learning fine-grained image-text knowledge.

4.4 Ablation study

To conduct an ablation study, we remove the fine-grained text encoder, the fine-grained image encoder, the fine-grained image-text matching predictor and the sense auto-complementing module. The results, as shown in Table 4, illustrate that our fine-grained text encoder and sense auto-complementing module are the most critical components. The fine-grained image encoder and the matching predictor can further contribute to the performance of understanding on fine-grained image-text knowledge.

5 Conclusion

This paper proposes a FCLL model for inference on fine-grained image-text knowledge and complementing additional English contextual information to phrases composed by limited words. Besides, we demonstrate that our fine-grained contrastive language-image learning approach and sense auto-complementing module support FCLL for understanding fine-grained image-text knowledge. Moreover, we construct a new multimodal-multilingual fine-grained image-text knowledge base, which can be applied to visual WSD to improve FCLL performance. The results on the benchmark test set from SemEval-2023 Task 1 show that FCLL achieves an average H@1 of 72.56% and an average MRR of 82.22%, ranking at the first in overall evaluation. Source codes and the knowledge base have been released to facilitate future fine-grained language-image research.

References

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.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

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.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten years of babelnet: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4559–4567. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2649.

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.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8758–8765.

Federico Scozzafava, Marco Maru, Fabrizio Brignone, Giovanni Torrisi, and Roberto Navigli. 2020. Personalized PageRank with syntagmatic information for multilingual word sense disambiguation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–46, Online. Association for Computational Linguistics.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Ming Wang and Yinglin Wang. 2020. A synset relation-enhanced framework with a try-again mechanism for word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6229–6240, Online. Association for Computational Linguistics.