words could not be determined unequivocally.

5 Participating Systems

This Section is devoted to the participating systems. First, we briefly describe the rules of the competition. Subsequently, we provide an overview of the data and approaches used by participants. Then, we focus on some of the best-scoring systems and provide a breakdown of the techniques adopted. We report the three best-performing teams for each sub-task and language combination in Tables 6 and 7. All results are publicly available on the official MCL-WiC page on GitHub[11]. For each winning team, we show only the best performance in the corresponding category.

5.1 Rules of the competition

Participants were given no constraints as far as data was concerned; for instance, the development data could be used for training or it was allowed to enrich the provided data by constructing new datasets in an automatic or semi-automatic fashion. Furthermore, we allowed more than one participant for each team. Participating teams could upload up to five submissions, each including up to 9 language combinations for the two sub-tasks.

5.2 Data

Multilingual sub-task As far as English is concerned, the majority of participating systems used the MCL-WiC training and development data. Some participants also used the data derived from WiC and XL-WiC. Furthermore, automatically-constructed WiC-like datasets were obtained by some participants, starting from semantic resources such as SemCor (Miller et al., 1993), WordNet and the Princeton WordNet Gloss Corpus (PWNG)[12], or by automatically translating available datasets into English. The available data was also enriched via sentence reversal augmentation (given a sentence pair, the two sentences were swapped). In some cases, the development and trial[13] data was used to enrich the training data.

As regards languages other than English, most participants used XL-WiC data, or new training and development datasets were obtained by splitting the MCL-WiC language-specific development data. Alternatively, in zero-shot scenarios, participants trained their models using the English training data. Furthermore, some participants augmented the training and development data by including the trial data. Also in this case, training and development splits were augmented via sentence reversal.

Cross-lingual sub-task In the cross-lingual sub-task, most participants used the MCL-WiC English training and development data in zero-shot settings. A smaller group of participants used WiC and XL-WiC data. Some participants created additional training and development data from other resources such as the Open Multilingual WordNet and PWNG. Additional training and development data was produced via Machine Translation.

5.3 Approaches

Multilingual sub-task Most participants used XLM-RoBERTa (Conneau et al., 2020) as pre-trained language model to obtain contextual representations of the target occurrences. Other models frequently used by participants were mBERT, RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), ELECTRA (Clark et al., 2019) and ERNIE (Sun et al., 2020). The majority of participants made use of fine-tuned contextualized embeddings and used logistic regression to perform binary classification. Some participants used ensembles and majority voting.

Cross-lingual sub-task Also in this sub-task, XLM-RoBERTa was the most used multilingual language model. Again, the majority of systems obtained contextualized embeddings, passing them to a logistic regression unit. In this case, participants mainly explored zero-shot approaches. Some participants made use of ensembles, adversarial training, pseudo-labelling (Wu and Prasad, 2017) and cross-validation techniques.

5.4 Competition and best-scoring systems

The MCL-WiC competition took place on the CodaLab[14] open Web-based platform and reported 170 participants, out of which 48 uploaded one or more datasets. Overall, 170 submissions were received, the majority of which were focused on the multilingual sub-task and specifically on the En-En dataset. As far as the evaluation metric was concerned, systems were tested using the accuracy

[11] https://github.com/SapienzaNLP/mcl-wic

[12] http://wordnetcode.princeton.edu/

[13] As trial data, we provided 4 instances for each sub-task and dataset.

[14] https://competitions.codalab.org/competitions/27054

Dataset	Team	Score
Ar-Ar	Cam	**84.8**
	LIORI	84.6
	MCL@IITK; DeathwingS	84.5
En-En	MCL@IITK; oyx	**93.3**
	zhestyatsky	92.7
	Cam	92.5
Fr-Fr	MCL@IITK	**87.5**
	Cam	86.5
	LIORI	86.4
Ru-Ru	Cam	**87.4**
	LIORI	86.6
	godzilla	86.5
Zh-Zh	stce	**91.0**
	godzilla	90.8
	PALI	90.5

Table 6: Multilingual section: five best-scoring systems by language combination.

score. In what follows, we provide insights regarding the approaches adopted by some of the best-performing participating systems, based on the information we received.

Cam The Cam team (Yuan and Strohmaier, 2021) made use of the WiC and XL-WiC datasets in addition to the MCL-WIC data. Furthermore, examples from the Sense Complexity Dataset (Strohmaier et al., 2020, SeCoDa) and the Cambridge Advanced Learner's Dictionary (CALD) were extracted. Cam used pre-trained XLM-RoBERTa as underlying language model and added two additional layers on top to perform binary classification with tanh and sigmoid activation, respectively. As input, the following items were concatenated: the representation corresponding to the first token of the sequence, the representations of the target words in both sentences, as well as the absolute difference, cosine similarity and pairwise distance between the two vectors. When the target word was split into multiple sub-tokens, Cam took the average representation rather than the first sub-token. Finally, a two-step training strategy was applied: 1) pre-training the system using out-of-domain data, i.e. WiC, XL-WiC, SeCoDa and CALD; 2) fine-tuning the system on MCL-WiC data.

godzilla godzilla enriched the MCL-WiC training data by automatically constructing a dataset starting from WordNet and using Machine Translation. Different types of pre-trained models, such as RoBERTa and XLM-RoBERTa, were adopted. godzilla highlighted the target words by surrounding them with special markings on both sides and appending the target words to the end of each sentence. As architecture, this system used the next sentence prediction models from the hugging face[15] library. Given the strong connection between En-Ar, En-Fr, En-Ru, En-Zh test datasets, pseudo-tagging was used for each language combination. Finally, godzilla applied label smoothing and model merging.

LIORI The LIORI[16] team (Davletov et al., 2021) used the datasets provided in the MCL-WiC competition. Specifically, the training data was enriched with 70% of the development data for Arabic, Chinese, French and Russian, and the whole trial data. Optionally, data augmentation was performed by swapping sentences in each example. LIORI fine-tuned XLM-RoBERTa on a binary classification task and used a 2-layered feed-forward neural network on top of the language model with dropout and the tanh activation function. Sentences in each pair were concatenated by the special token "</s>" and fed to XLM-RoBERTa. As input, the model took the concatenation of the contextualized embeddings of the target words, aggregating over sub-tokens either by max pooling, or just by taking the first sub-token. LIORI used a voting ensemble composed of three models: the first model trained with data augmentation, using the concatenations of the first sub-tokens of the target words; the second trained with data augmentation using max-pooling over sub-tokens; finally, the third trained without data augmentation and using concatenations of the first sub-tokens.

stce stce used the MCL-WiC datasets and built additional training data using HowNet (Dong and Dong, 2003). Furthermore, the training data was enriched by pseudo-labelling the test datasets. Data cleaning was performed and target words were surrounded by special markings. The main language model used was XLM-RoBERTa-large. During the training process, dynamic negative sampling was performed for each batch of data fed to the model. At the same time, stce adopted the Fast Gradient Method and added disturbance to the embedding layer to obtain more stable word representations.

[15] https://huggingface.co/

[16] The following member of the team LIORI took part in the competition: davletov.

Dataset	Team	Score
En-Ar	PALI	**89.1**
	godzilla	87.0
	Cam; LIORI	86.5
En-Fr	PALI	**89.1**
	godzilla	87.6
	LIORI	87.2
En-Ru	PALI	**89.4**
	godzilla	88.5
	RyanStark; rxy1212	87.3
En-Zh	PALI; RyanStark	**91.2**
	Cam	88.8
	MagicPai	88.6

Table 7: Cross-lingual sub-task: three best-scoring systems by language combination.

zhestyatsky Zhestiankin and Ponomareva (2021) augmented the English MCL-WiC training and development data with WiC. Training and development data were split randomly to create a larger training sample which included 97.5% of the data, while leaving only 2.5% for the new development dataset. Then, bert-large-cased embeddings were fine-tuned using AdamW as optimizer with a learning rate equal to 1e-5. Each sentence was split by BertTokenizerFast into 118 tokens maximum. The model was trained for 4.5 epochs and stopped by Early Stopping with patience equal to 2. For each sentence, zhestyatsky took the embeddings of all sub-tokens corresponding to the target word and max pooled them into one embedding. Subsequently, zhestyatsky evaluated the cosine similarity of these embeddings and activated this value through ReLU.

MCL@IITK First, the MCL@IITK[17] team (Gupta et al., 2021) pre-processed the sentences by adding a signal, either double quotes on both sides of the target word, or the target word itself appended to the end of the sentence. For En-En, MCL@IITK enriched the MCL-WiC training data using sentence reversal augmentation, WiC and SemCor. MCL@IITK obtained embeddings of the target words using the last hidden layer, and passed them to a logistic regression unit. MCL@IITK used ELECTRA, ALBERT, and XLM-RoBERTa as language models and submitted probability sum ensembles. For the non-English multilingual subtask, MCL@IITK used XLM-RoBERTa only and tackled all four language pairs jointly. A 9:1 train-dev split with sentence reversal augmentation was used on the non-English dev data, in addition to En-En train data and XL-WiC with an ensemble model. For the cross-lingual subtask, ELECTRA embeddings were used. The models were trained on partly back-translated En-En train set and validated on back-translated En-En development set.

PALI The PALI[18] team (Xie et al., 2021) enriched the MCL-WiC data using WordNet while keeping the original cross-lingual data to maintain the target words in the cross-lingual data. After text pre-processing, task-adaptive pre-training was performed using the MCL-WiC data. The target words were surrounded by special symbols. PALI used XLM-RoBERTa as main language model and took its final output layer, concatenating the [CLS] token with the embeddings of the target occurrences in each sentence pair. To increase the training data, PALI exchanged the order of 20% of the sentence pairs. During training, lookahead (AdamW) was used together with adversarial training implemented by the Fast Gradient Method to obtain more stable word representations. Hyperparameters were tuned through trial-and-errors. The models of stratified 5-fold cross-validation were averaged to yield the final prediction results.

6 Baselines

Following Raganato et al. (2020), we used a baseline transformer-based binary classifier. Thus, first, given a sentence pair, a dense representation is obtained for each target occurrence. As indicated in Devlin et al. (2019), in the case that a target occurrence is split into multiple sub-tokens, the first sub-token is selected. The resulting representations are then given as input to a binary classifier implemented following Wang et al. (2019). We selected the Adam optimizer (Kingma and Ba, 2015) with learning rate and weight decay equal to 1e-5 and 0, respectively, and trained for 10 epochs.

We experimented with two different contextualized embedding models: BERT (base-multilingual-cased) and XLM-RoBERTa (base). As for the data, in contrast to most participants, we made use of the data provided for the task only. We used En-En as training and development data for English. As for other language combinations, we trained on En-En and validated both on En-En or and on the other

[17]The following members of the MCL@IITK team took part in the competition: jaymundra, rohangpt and dipakam.

[18]The following members of the PALI team took part in the competition: endworld and xsysigma.