Model	Ar-Ar	En-En	Fr-Fr	Ru-Ru	Zh-Zh	En-Ar	En-Fr	En-Ru	En-Zh
mBERT$_1$	**76.2**	84.0	**78.7**	74.5	77.5	65.9	71.6	68.2	**68.9**
XLMR-base$_1$	75.4	**86.6**	77.9	**76.5**	**78.5**	**67.7**	**71.8**	**74.2**	66.1
mBERT$_2$	**76.4**	84.0	**78.7**	74.6	76.6	62.0	69.4	66.7	64.2
XLMR-base$_2$	75.4	**86.6**	77.7	**76.5**	**78.9**	**67.7**	**74.9**	**74.2**	**71.3**

Table 8: Accuracy of baselines for multilingual and cross-lingual sub-tasks. Columns indicate the test set used. In setting 1, we used the En-En training data and the En-En development data. In setting 2, we used the En-En training data and the corresponding development datasets in languages other than English.

language multilingual development data. Table 8 reports the best training results according to the corresponding validation.

7 Results and Discussion

In this section, we discuss the results achieved in our competition. Overall, the MCL-WiC dataset allows systems to attain high performances, in the 85-93% accuracy range. This leads us to hypothesize that, in general, systems were able to develop a good ability in capturing sense distinctions without relying on a fixed sense inventory.

When compared to the proposed baselines, we observe that best-performing systems were able to achieve an absolute improvement of up to 27.1 points over the corresponding baselines (e.g. on En-Ar, cf. Tables 7 and 8). Both our baselines and the systems developed by participants confirm that, in this task, XLM-RoBERTa outperforms BERT in most language combinations. The highest score was obtained in En-En, with the best system achieving 93.3% accuracy. Note that our baselines were also able to attain good performances in En-En, i.e. 84.0% using BERT and 86.6% with XLM-RoBERTa, without benefiting from additional training and development data. Interestingly, Chinese was the language which achieved the second-best results, both in Zh-Zh and En-Zh, attaining on average results which were considerably higher. Instead, Arabic seems to have been the most difficult language for participants, especially in Ar-Ar. A reason for this result, deserving further exploration, could lie in morpho-semantic features inherent in Arabic, which we briefly outlined in Section 4.

Zero-shot approaches differ in the performances achieved by participants in the two sub-tasks: in the cross-lingual sub-task participants were able to achieve slightly better performances than those in the multilingual setting, most probably thanks to the presence of English in both the training and the test data, and, more in general, to the availability of English WiC-style datasets which could be used to enrich the already provided data. With the exception of Chinese, instead, on the multilingual sub-task we observe a performance drop between 1.6 and 4.3%.

Finally, we note that performance boosts were observed across the board when using data augmentation, especially by swapping the two sentences within a pair or by coupling the second sentences of two pairs sharing the same first sentence and the same meaning. Another consistent performance increase, observed both in the multilingual and in the cross-lingual sub-task, was obtained when adding a signal on both sides of the target occurrences.

8 Conclusions

In this paper, we described the SemEval-2021 Task 2 and introduced Multilingual and Cross-lingual Word-in-Context (MCL-WiC), the first entirely manually-curated WiC-style dataset in five European and non-European languages, namely Arabic, Chinese, English, French and Russian. MCL-WiC allows the inherent ability of systems to discriminate between word senses within the same language to be tested, and also, interestingly, within cross-lingual scenarios in which a system is evaluated in two languages at the same time, namely English and one of the remaining MCL-WiC languages.

While current Word-in-Context datasets focus primarily on single tokens, as a suggestion for future work we would like to further explore the integration of multi-word expressions and idiomatic phrases into a Word-in-Context task. This would allow us to investigate the intrinsic ability of a system to correctly discriminate the semantics of such linguistic constructs, especially those whose meaning is not compositional, i.e. it cannot be derived by combining the meaning of each of their individual components.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 and the ELEXIS project No. 731015 under the European Union's Horizon 2020 research and innovation programme.

We gratefully thank Luisa Borchio, Ibraam Abdelsayed, Anna Guseva, Zhihao Lyu and Beatrice Buselli for their valuable annotation work.

References

Carlos Santos Armendariz, Matthew Purver, Senja Pollak, Nikola Ljubešić, Matej Ulčar, Ivan Vulić, and Mohammad Taher Pilehvar. 2020. SemEval-2020 Task 3: Graded Word Similarity in Context. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 36–49.

Carlos Santos Armendariz, Matthew Purver, Matej Ulcar, Senja Pollak, Nikola Ljubesic, Marko Robnik-Sikonja, Mark Granroth-Wilding, and Kristiina Vaik. CoSimLex: A resource for evaluating graded word similarity in context. In *Proceedings of The 12th Lan-guage Resources and Evaluation Conference*, page 5878–5886.

Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. 2020. EViLBERT: Learning task-agnostic multimodal sense embeddings. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 481–487. International Joint Conferences on Artificial Intelligence Organization.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Adis Davletov, Nikolay Arefyev, Denis Gordeev, and Alexey Rey. 2021. LIORI at SemEval-2021 Task 2: Span Prediction and Binary Classification approaches to Word-in-Context Disambiguation. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*, Bangkok, Thailand (online). Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Zhendong Dong and Qiang Dong. 2003. HowNet-a hybrid language and knowledge resource. In *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*, pages 820–824. IEEE.

Rohan Gupta, Jay Mundra, Deepak Mahajan, and Ashutosh Modi. 2021. MCL@IITK at SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation using Augmented Data, Signals, and Transformers. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation*, Bangkok, Thailand. Association for Computational Linguistics.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 95–105.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5682–5691. Association for Computational Linguistics.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 Task 10: English lexical substitution task. In *Proceedings of the 4th International Workshopon Semantic Evaluations (SemEval-2007)*, page 48–53.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. Context2Vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 51–61. ACL.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 Task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 9–14.

George A. Miller, R.T. Beckwith, Christiane D. Fellbaum, D. Gross, and K. Miller. 1990. Introduction to WordNet: an online lexical database. *International Journal of Lexicography*, 3(4).

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

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.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2015. Efficient non-parametric estimation of multiple embeddings per word in vector space. *CoRR*, abs/1504.06654:1059–1069.

Maria Pelevina, Nikolay Arefiev, Chris Biemann, and Alexander Panchenko. 2016. Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP, Rep4NLP@ACL 2016, Berlin, Germany, August 11, 2016*, pages 174–183. Association for Computational Linguistics.

Mohammad Taher Pilehvar and José Camacho-Collados. 2019. WiC: the Word-in-Context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1267–1273. Association for Computational Linguistics.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. Xl-wic: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020a. SensEmBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proc. of AAAI*, pages 8758–8765.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020b. With More Contexts Comes Better Performance: Contextualized Sense Embeddings for All-Round Word Sense Disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, page 3528–3539. Association for Computational Linguistics.

David Strohmaier, Sian Gooding, Shiva Taslimipoor, and Ekaterina Kochmar. 2020. SeCoDa: Sense complexity dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5962–5967, Marseille, France. European Language Resources Association.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 3261–3275.

Hao Wu and Saurabh Prasad. 2017. Semi-supervised deep learning using pseudo labels for hyperspectral image classification. *IEEE Transactions on Image Processing*, 27(3):1259–1270.

Shuyi Xie, Jian Ma, Haiqin Yang, Lianxin Jiang, Yang Mo, and Jianping Shen. 2021. PALI at SemEval-2021 task 2: Fine-Tune XLM-RoBERTa for Word in Context Disambiguation. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*.

Zheng Yuan and David Strohmaier. 2021. Cambridge at SemEval-2021 Task 2: Neural WiC-Model with Data Augmentation and Exploration of Representation. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*.

Boris Zhestiankin and Maria Ponomareva. 2021. Zhestyatsky at SemEval-2021 Task 2: ReLU over Cosine Similarity for BERT Fine-tuning. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*.