Cambridge at SemEval-2021 Task 2: Neural WiC-Model with Data Augmentation and Exploration of Representation

Zheng Yuan and **David Strohmaier**
ALTA Institute, University of Cambridge, United Kingdom
Department of Computer Science and Technology, University of Cambridge, United Kingdom
{zheng.yuan, david.strohmaier}@cl.cam.ac.uk

Abstract

This paper describes the system of the Cambridge team submitted to the SemEval-2021 shared task on Multilingual and Cross-lingual Word-in-Context Disambiguation. Building on top of a pre-trained masked language model, our system is first pre-trained on out-of-domain data, and then fine-tuned on in-domain data. We demonstrate the effectiveness of the proposed two-step training strategy and the benefits of data augmentation from both existing examples and new resources. We further investigate different representations and show that the addition of distance-based features is helpful in the word-in-context disambiguation task. Our system yields highly competitive results in the cross-lingual track without training on any cross-lingual data; and achieves state-of-the-art results in the multilingual track, ranking first in two languages (Arabic and Russian) and second in French out of 171 submitted systems.

1 Introduction

Polysemy still poses a great challenge to natural language processing (NLP) applications. Depending on its context, an ambiguous word can refer to multiple, potentially unrelated, meanings. Recently, as an application of Word Sense Disambiguation (WSD) (Navigli, 2009, 2012), Word-in-Context (WiC) disambiguation has been framed as a binary classification task to identify if the occurrences of a target word in two contexts correspond to the same meaning or not. The release of the WiC dataset (Pilehvar and Camacho-Collados, 2019), followed by the Multilingual Word-in-Context (XL-WiC) dataset (Raganato et al., 2020), has helped provide a common ground for evaluating and comparing systems while encouraging research in WSD and context-sensitive word embeddings.

In this paper, we describe our submission to the SemEval-2021 shared task on Multilingual and Cross-lingual Word-in-Context (MCL-WiC) Disambiguation (Martelli et al., 2021), which involves determining whether a word shared by two sentences in the same language (multilingual track) or across different languages (cross-lingual track) has the same meaning in both contexts. Compared to previous WiC and XL-WiC benchmarks, two new languages are introduced as well as a cross-lingual track where systems are evaluated under a 'zero-shot' setting.

The MCL-WiC task directly classifies pairs of sentences with regard to the meaning of the shared word. By turning WSD into a binary comparison task, MCL-WiC avoids the need for sense tags of previous WSD shared tasks (Manandhar et al., 2010; Navigli et al., 2013; Moro and Navigli, 2015). It also resembles the Word Sense Alignment (WSA) task (Ahmadi et al., 2020) more closely, in which definitions from different dictionaries have to be aligned. Contextualised word embeddings and pre-trained Transformer-based (Vaswani et al., 2017) language models have been increasingly applied to these tasks and state-of-the-art results have been reported (Hadiwinoto et al., 2019; Vial et al., 2019; Levine et al., 2020; Raganato et al., 2020; Pais et al., 2020; Manna et al., 2020; Lenka and Seung-Bin, 2020).

In line with previous research, we develop a neural system based on pre-trained multilingual masked language model XLM-R (Conneau et al., 2020). Additionally, we introduce three distance-based features to be used together with the widely used sequence and token representations for MCL-WiC disambiguation. To further improve system performance, we apply automatic data augmentation and extract examples from multiple external resources. A two-step training strategy is then employed to make use of both in-domain and out-of-

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 730–737
Bangkok, Thailand (online), August 5–6, 2021.

Split	Multilingual					Cross-lingual			
	EN-EN	AR-AR	FR-FR	RU-RU	ZH-ZH	EN-AR	EN-FR	EN-RU	EN-ZH
Train	8,000	-	-	-	-	-	-	-	-
Dev	1,000	1,000	1,000	1,000	1,000	-	-	-	-
Test	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Table 1: Number of instances in the official MCL-WiC dataset in each multilingual and cross-lingual sub-track.

Example	Instance	Sentence #1	Sentence #2	Label
(a)	Existing	There is never any point in trying to make oneself heard over **noise**.	We have formulated a programme to address the traffic **noise** impact of existing roads.	T
	Existing	There is never any point in trying to make oneself heard over **noise**.	He went to bed but could not fall asleep because of the **noise**.	T
	New	We have formulated a programme to address the traffic **noise** impact of existing roads.	He went to bed but could not fall asleep because of the **noise**.	T
(b)	Existing	Wages have declined sharply as a proportion of the **subsistence** minimum.	Agriculture, largely of a **subsistence** nature, is the main economic activity.	T
	Existing	Wages have declined sharply as a proportion of the **subsistence** minimum.	The third member of the Committee is paid a daily fee for each working day plus **subsistence** allowance.	F
	New	Agriculture, largely of a **subsistence** nature, is the main economic activity.	The third member of the Committee is paid a daily fee for each working day plus **subsistence** allowance.	F

Table 2: Sample sentence-pair instances selected for data augmentation. The target words are marked in bold. Examples are extracted from the MCL-WiC dataset.

domain[1] data.

In the remainder of the paper, we present the MCL-WiC disambiguation shared task in Section 2 and our approach in Section 3. In Section 4, we describe the experiments and present results on the development set. Section 5 summarises the official evaluation results. Finally, we provide an analysis of our system in Section 6 and conclude in Section 7.

2 Task Description

The MCL-WiC dataset used in the shared task consists of sentence pairs sharing the same target word in the same language or across different languages. The task considers five languages: Arabic (AR), Chinese (ZH), English (EN), French (FR) and Russian (RU); and contains five multilingual (EN-EN, AR-AR, FR-FR, RU-RU, ZH-ZH)[2] and four cross-lingual (EN-AR, EN-FR, EN-RU, EN-ZH) sub-tracks. Training data is available for the multilingual EN-EN sub-track only, and development data is available for all five multilingual sub-tracks. No cross-lingual training or development data is provided. Statistics of the MCL-WiC dataset are presented in Table 1.

Results are computed using the accuracy measure, i.e. the ratio of correctly predicted instances (true positives or true negatives) to the total number of instances.

3 Approach

3.1 Data augmentation

Each instance in the (*)WiC datasets (i.e. WiC, XL-WiC and MCL-WiC) is composed of a target word and two sentences in which the target word occurs. We notice that there are cases where the same sentence appears in multiple instances. As shown in Table 2, two existing instances, which share the same target word, contain the same first sentence, but different second sentences. Therefore, we construct new instances by pairing the second sentences from these existing instances and assign labels based on the original labels:

- If both existing instances are positive ('T', i.e. the target word is used in the same meaning), the resulting instance should be positive ('T') as well - see Example (a) in Table 2:

$$M(w_{s_1}) = M(w_{s_2}), M(w_{s_1}) = M(w_{s_3}) \Rightarrow M(w_{s_2}) = M(w_{s_3})$$

[1] In this paper, we use the term 'out-of-domain' to refer to data from additional resources (i.e. not provided by the shared task organisers) - see Section 4.1 for more details.

[2] [language of sentence #1]-[language of sentence #2]

where $M(w_{s_n})$ indicates the meaning of the target word (w) used in sentence n (s_n).

- If one of them is positive ('T') and the other is negative ('F', i.e. the target word is used in a different meaning), the new instance should then be negative ('F') - see Example (b) in Table 2:

$$M(w_{s_1}) = M(w_{s_2}), M(w_{s_1}) \neq M(w_{s_3}) \Rightarrow M(w_{s_2}) \neq M(w_{s_3})$$

3.2 Model

Following Raganato et al. (2020), we use pre-trained XLM-R as the underlying language model, which is a Transformer-based multilingual masked language model that has been trained on one hundred languages (Conneau et al., 2020). Unlike previous WiC and XL-WiC models employing a logistic regression classifier (Wang et al., 2019; Raganato et al., 2020), we add two additional layers on top of the Transformer-based model to perform classification: a linear layer with $tanh$ activation, followed by another linear layer with $sigmoid$ activation.

The model takes as input the two sentences in each instance. For the representation to be fed into the linear layers, we concatenate the representation corresponding to the first special token ($[s]$) of the input sequence,[3] the vector representations of the target word in the first ($[w_{s_1}]$) and second sentences ($[w_{s_2}]$), as well as the element-wise absolute difference, cosine similarity (cos) and Euclidean distance ($dist$) between these two vectors:

$$[s; w_{s_1}; w_{s_2}; |w_{s_1} - w_{s_2}|; cos; dist] \quad (1)$$

For those cases where the target word is split into multiple sub-tokens, we take the averaged representation rather than the representation of its first sub-token, which has been used in previous work (Wang et al., 2019; Raganato et al., 2020).[4]

3.3 Training strategy

Inspired by the success of multi-stage training for tasks like grammatical error correction (Kiyono et al., 2019; Omelianchuk et al., 2020; Yuan and Bryant, 2021) and machine translation (Zhang et al., 2018), we employ a two-step training strategy: 1) pre-training on out-of-domain data; and 2) fine-tuning with in-domain MCL-WiC data.

[3]The [s] token in XLM-R is equivalent to the [CLS] token in BERT (Devlin et al., 2019).

[4]Our preliminary experiments show that using the averaged representation is more effective than that of the first sub-token.

Sentence #1	I went outside to get some fresh **air**.	(A2)
Sentence #2	He has an **air** of authority.	(C2)
Label	F	

Table 3: A sentence-pair example extracted from CALD. The target words are marked in bold. A2: elementary English, C2: proficiency English.

4 Experiments

4.1 Data

In addition to the MCL-WiC dataset provided by the shared task organisers, we introduce two types of out-of-domain data: 1) (*)WiC datasets: WiC and XL-WiC; and 2) sentence pairs constructed with examples extracted from datasets that have been annotated with both complexity and sense information: the Cambridge Advanced Learner's Dictionary (CALD)[5] and SeCoDa (Strohmaier et al., 2020).

WiC The English WiC dataset was created using example sentences from WordNet (Fellbaum, 1998), VerbNet (Kipper-Schuler, 2005), and Wiktionary. We extract 6,066 labelled instances (by removing those without gold labels) and use them for the shared task.

XL-WiC The XL-WiC dataset extends the WiC dataset to 12 more languages from two resources: multilingual WordNet for ZH, Bulgarian (BG), Croatian (HR), Danish (DA), Dutch (NL), Estonian (ET), Farsi (FA), Japanese (JA) and Korean (KO); and multilingual Wiktionary for FR, German (DE), Italian (IT). In total, the XL-WiC dataset contains 112,430 labelled non-English sentence pairs, including 3,046 ZH-ZH instances and 48,106 FR-FR ones.[6] In contrast to the MCL-WiC task, the XL-WiC dataset does not include any cross-lingual sentence pairs.

CALD The CALD contains information about which words and which meanings of those words are known and used by learners at each Common European Framework of Reference (CEFR) level from A1 (beginner) to C2 (proficiency English). Only example sentences sharing the same target word, that is used in a different meaning as well as at a different CEFR level, are paired. In this

[5]https://www.englishprofile.org/wordlists/evp

[6]In the XL-WiC dataset, the number of instances varies considerably by language - see Raganato et al. (2020) for more details.

Dataset	Language	#instances (w/o aug.)	#instances (with aug.)	Training stage
WiC	EN-EN	6,066	8,276	pre-training
XL-WiC(FR)	FR-FR	48,016	54,771	pre-training
XL-WiC(ZH)	ZH-ZH	3,046	3,370	pre-training
XL-WiC	All*	112,430	127,765	pre-training
CALD	EN-EN	34,205	-	pre-training
SeCoDa	EN-EN	10,712	-	pre-training
MCL-WiC-train	EN-EN	8,000	10,798	fine-tuning
MCL-WiC-dev(EN)	EN-EN	1,000	-	development

Table 4: Summary of datasets used in our experiments. All*: FR-FR, ZH-ZH, BG-BG, HR-HR, DA-DA, NL-NL, ET-ET, FA-FA, JA-JA, KO-KO, DE-DE, and IT-IT; w/o aug.: without data augmentation; with aug.: with data augmentation.

#	Pre-training data	EN-EN	AR-AR	FR-FR	RU-RU	ZH-ZH	Avg.
1	$\text{WiC}_{\text{aug.}}$ + $\text{XL-WiC(FR\&ZH)}_{\text{aug.}}$	89.90	78.10	80.10	83.60	78.20	81.98
2	$\text{WiC}_{\text{aug.}}$ + $\text{XL-WiC}_{\text{aug.}}$	89.60	79.20	82.90	85.40	**79.30**	83.28
3	$\text{WiC}_{\text{aug.}}$ + $\text{XL-WiC}_{\text{aug.}}$ + CALD + SeCoDa	90.30	**80.20**	**83.30**	**86.30**	76.90	83.40
4	Ensemble (MV)	**90.80**	79.70	83.00	85.40	**79.30**	**83.64**

Table 5: Performance of individual systems and the ensemble on MCL-WiC-dev (multilingual track). The best results in each sub-track are marked in bold. Avg.: averaged accuracy.

way, sentence pairs are encoded with additional word complexity information. In total, we generate 34,205 negative EN-EN instances.[7] An example is given in Table 3.

SeCoDa is an English language corpus of words annotated with both complexity and word senses. The original data comes from three sources: professional News, WikiNews and Wikipedia articles. The senses are drawn from the CALD and come at two levels of granularity. To use this dataset for the MCL-WiC task, sentences sharing an annotated word are paired: if the word shares a sense, the pair of sentences is labelled as 'T'; otherwise, it is labelled as 'F'. We use the finer level of granularity for this assignment. Overall, we extract 10,712 labelled pairs (9,015 positive and 1,697 negative).

All the data introduced in this section is regarded as out-of-domain data and therefore used in the pre-training stage, and the in-domain MCL-WiC training data is used in the fine-tuning stage. For development, we use only the EN-EN MCL-WiC development set (**MCL-WiC-dev(EN)** - see Table 4). A single model is developed to target all multilingual and cross-lingual tracks. It is worth noting that neither the multilingual AR-AR and RU-RU data nor the cross-lingual data is used for training, i.e. zero-shot.

The data augmentation method proposed in Section 3.1 is applied to the (*)WiC datasets, but not to CALD or SeCoDa. Detailed statistics of the corpora used in our experiments are presented in Table 4.

4.2 Training details

In our experiments, models are trained by minimising the binary cross-entropy loss between their prediction and the gold label. We use the AdamW optimiser (Loshchilov and Hutter, 2019) with a fixed learning rate of 1e-5 for all models.[8] We use a dropout layer with a dropout probability of 0.2. The input texts are processed in batches of 8 and are padded or truncated to a length of 182.[9] We select the model with the best validation accuracy on **MCL-WiC-dev(EN)**. Each model is trained on one NVIDIA Tesla P100 GPU.

4.3 Results

We construct three pre-training sets using different combinations of the out-of-domain data and

[7] Due to time limitations, we have not used any positive instances from CALD and leave it for future work.

[8] We use the pre-trained `xlm-roberta-large` model provided by Hugging Face (https://huggingface.co/) transformers library (Wolf et al., 2020).

[9] We use a maximum sequence length of 182, which is the length of the longest input sequence in the MCL-WiC training set.

System	EN-EN		AR-AR		FR-FR		RU-RU		ZH-ZH	
	Acc.	Rank	Acc.	Rank	Acc.	Rank	Acc.	Rank	Acc.	Rank
#4	92.50	6	84.80	1	86.50	2	87.40	1	85.80	13

Table 6: Official results of our best submitted system on the MCL-WiC test set (multilingual track). Acc.: accuracy.

System	EN-AR		EN-FR		EN-RU		EN-ZH	
	Acc.	Rank	Acc.	Rank	Acc.	Rank	Acc.	Rank
#1	86.50	6	85.70	10	86.80	9	88.80	5

Table 7: Official results of our best submitted system on the MCL-WiC test set (cross-lingual track).

train three systems. All these systems are fine-tuned with the augmented version of the MCL-WiC training set (**MCL-WiC-train$_{\text{aug.}}$** - see Table 4). Results on the MCL-WiC multilingual development set are presented in Table 5, where **WiC$_{\text{aug.}}$** is the augmented version of the WiC dataset, **XL-WiC$_{\text{aug.}}$** is the augmented version of the full XL-WiC dataset for all 12 languages, and **XL-WiC(FR&ZH)$_{\text{aug.}}$** is a subset of **XL-WiC$_{\text{aug.}}$**, containing only examples in FR and ZH (i.e. the only two languages shared by MCL-WiC and XL-WiC).

We can see that adding pre-training examples in other languages from the XL-WiC dataset improves the results for all languages except for EN-EN, where the accuracy slightly drops from 89.90 to 89.60 (Table 5 #2). Interestingly, Raganato et al. (2020) also reported performance gains in all languages after adding multilingual data. Examples from different languages can still help models better generalise across languages. The addition of English data from CALD and SeCoDa is also beneficial, yielding further improvements in all languages except for ZH-ZH (#3). Finally, the predictions from all three systems are used in an ensemble model. For each final prediction, the majority vote (MV) of these predictions is taken, i.e. the prediction with the most votes is chosen as final prediction. The ensemble model yields the highest averaged score, as well as in EN-EN and ZH-ZH sub-tracks (#4), suggesting that all three systems (#1, #2 and #3) are complementary.

5 Official evaluation results

We submit our systems to all multilingual and cross-lingual tracks. The official results of our best submission for each track are reported in Table 6 and Table 7. Our ensemble system (**System #4**) achieves state of the art in the multilingual track, ranking first in both AR-AR and RU-RU sub-tracks without any AR or RU training data, and second in FR-FR out of 171 submitted systems. For the cross-lingual track, our 'zero-shot' system (**System #1**) is consistently within the top ten ranks out of 171 total submissions.[10]

6 Analysis

6.1 Effect of two-step training

We propose a two-step training strategy to make use of both in-domain and out-of-domain data. To investigate the effectiveness of both training steps, we undertake an ablation study, in which we remove one training step at a time. Table 8 presents the ablation test results of the system pre-trained on **WiC$_{\text{aug.}}$ + XL-WiC(FR&ZH)$_{\text{aug.}}$**, and fine-tuned on **MCL-WiC-train$_{\text{aug.}}$** (i.e. **System #1** in Table 5).

The results of the ablation study demonstrate the effectiveness of the two-step training strategy, and show that it is crucial to have both pre-training and fine-tuning stages. Performance drops in all multilingual sub-tracks when removing either step, except for removing the pre-training step in FR-FR ($+0.70$). This is interesting as the model is pre-trained on data for only three sub-tracks, where FR-FR is one of them. For the other two languages, we observe the biggest performance decrease: ZH-ZH (-2.80) and EN-EN (-2.30). Overall, the fine-tuning stage seems more effective than the pre-training stage, though more data is used in the latter (10,798 for fine-tuning vs. 66,417 for pre-training), demonstrating the importance of having high-quality in-domain data.

[10] It is to be noted that the calculation of the rank counts every submission, even if they were made by the same team.

Ablated stage	EN-EN		AR-AR		FR-FR		RU-RU		ZH-ZH		Avg.	
	Acc.	Δ	Acc.	Δ	Acc.	Δ	Acc.	Δ	Acc.	Δ	Acc.	Δ
System #1	89.90	-	78.10	-	80.10	-	83.60	-	78.20	-	81.98	-
no pre-training	87.60	-2.30	77.10	-1.00	80.80	+0.70	82.70	-0.90	75.40	-2.80	80.72	-1.26
no fine-tuning	87.10	-2.80	75.80	-2.30	74.80	-5.30	79.60	-4.00	75.90	-2.30	78.64	-3.34

Table 8: Ablation test results of **System #1** on MCL-WiC-dev (multilingual track). Δ denotes the difference in accuracy (Acc.) score with respect to **System #1**.

Representation	EN-EN	AR-AR	FR-FR	RU-RU	ZH-ZH	Avg.
$[s; w_{s_1}; w_{s_2}]$	84.70	**77.30**	80.10	**83.80**	67.30	78.64
$[s; w_{s_1}; w_{s_2}; \lvert w_{s_1} - w_{s_2}\rvert]$	85.40	77.00	80.60	81.50	70.30	78.96
$[s; w_{s_1}; w_{s_2}; \lvert w_{s_1} - w_{s_2}\rvert; cos; dist]$	**87.60**	77.10	**80.80**	82.70	**75.40**	**80.72**

Table 9: Results of models using different representations. Systems are trained on MCL-WiC-train$_{\text{aug.}}$ and evaluated on MCL-WiC-dev (multilingual track). The best results in each sub-track are marked in bold. cos: cosine similarity, $dist$: Euclidean distance.

6.2 Comparison of representations

In our system, the representation pooled out from the underlying pre-trained language model is a combination of three vector representations (of the first token and target word in both sentences), and three distance-based features: the element-wise absolute difference, cosine similarity and Euclidean distance between the target word in both sentences (see Section 3.2). We further experiment with different representations and present our results in Table 9. We can see that our proposed representation yields the overall best performance across different languages, suggesting that the addition of all three distance-based features is indeed helpful.

7 Conclusion

In this paper, we presented the contribution of the Cambridge University team to the SemEval 2021 shared task on MCL-WiC Disambiguation. Using XLM-R, a pre-trained multilingual Transformer-based language model, as a starting point, we investigated automatic data augmentation, the use of multiple external datasets, multi-stage training strategies, and the representation of tokens and their distance. Our detailed analysis demonstrated the effectiveness of the two-step training strategy for making use of both in-domain and out-of-domain data, as well as the benefits of adding distance-based features to the representation for WiC disambiguation. Our best system yields highly competitive results in the cross-lingual track and achieves state-of-the-art results in the multilingual track, ranking first in two languages (AR and RU) and second in FR out of 171 total submissions.

Acknowledgments

We would like to thank Mariano Felice and Shiva Taslimipoor for their valuable comments and suggestions. This paper reports on research supported by Cambridge Assessment, University of Cambridge. This work was performed using resources provided by the Cambridge Service for Data Driven Discovery operated by the University of Cambridge Research Computing Service, provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/P020259/1), and DiRAC funding from the Science and Technology Facilities Council. We acknowledge NVIDIA for an Academic Hardware Grant.

References

Sina Ahmadi, John P. McCrae, Sanni Nimb, Fahad Khan, Monica Monachini, Bolette S. Pedersen, Thierry Declerck, Tanja Wissik, Andrea Bellandi, Irene Pisani, Thomas Troelsgård, Sussi Olsen, Simon Krek, Veronika Lipp, Tamás Váradi, László Simon, András Győrffy, Carole Tiberius, Tanneke Schoonheim, Yifat Ben Moshe, Maya Rudich, Raya Abu Ahmad, Dorielle Lonke, Kira Kovalenko, Margit Langemets, Jelena Kallas, Oksana Dereza, Theodorus Fransen, David Cillessen, David Lindemann, Mikel Alonso, Ana Salgado, José Luis Sancho, Rafael-J. Ureña-Ruiz, Kiril Simov, Petya Osenova, Zara Kancheva, Ivaylo Radev, Ranka Stanković, Andrej Perdih, and Dejan Gabrovšek. 2020. A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment. In *Proceedings of the 12th Language Resource and Evaluation Conference (LREC 2020)*, Marseille, France.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal,

Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard 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, 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. 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 5297–5306, Hong Kong, China. Association for Computational Linguistics.

Karin Kipper-Schuler. 2005. *Verbnet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania, USA. AAI3179808.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. 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 1236–1242, Hong Kong, China. Association for Computational Linguistics.

Bajčetíc Lenka and Yim Seung-Bin. 2020. Implementation of supervised training approaches for monolingual word sense alignment: Acdh-ch system description for the mwsa shared task at globalex 2020. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, pages 84–91.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR 2019)*.

Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. 2010. SemEval-2010 task 14: Word sense induction &disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68, Uppsala, Sweden. Association for Computational Linguistics.

Raffaele Manna, Giulia Speranza, Maria Pia di Buono, and Johanna Monti. 2020. Unior nlp at mwsa task - globalex 2020:siamese lstm with attention for word sense alignment. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, pages 76–83.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). In *Proceedings of the Fifteenth International Workshop on Semantic Evaluation (SemEval-2021)*.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado. Association for Computational Linguistics.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

Roberto Navigli. 2012. A quick tour of word sense disambiguation, induction and related approaches. In *SOFSEM 2012: Theory and Practice of Computer Science*, Lecture Notes in Computer Science, page 115–129. Springer.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

Vasile Pais, Dan Tufiș, and Radu Ion. 2020. MWSA task at GlobaLex 2020: RACAI's word sense alignment system using a similarity measurement of dictionary definitions. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 69–75, Marseille, France. European Language Resources Association.

Mohammad Taher Pilehvar and Jose 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, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. 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, Online. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Loic Vial, Benjamin Lecouteux, and Didier Schwab. 2019. Sense vocabulary compression through the semantic knowledge of wordnet for neural word sense disambiguation. *ArXiv*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32, pages 3266–3280. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zheng Yuan and Christopher Bryant. 2021. Document-level grammatical error correction. In *Proceedings of the Sixteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.