GlossReader at SemEval-2021 Task 2: Reading Definitions Improves Contextualized Word Embeddings

Maxim Rachinskiy[◇] **and Nikolay Arefyev**[△▽◇]
[◇]HSE University, Moscow, Russia
[△]Samsung Research Center Russia, Moscow, Russia
[▽]Lomonosov Moscow State University, Moscow, Russia
myurachinskiy@edu.hse.ru narefjev@cs.msu.ru

Abstract

Consulting a dictionary or a glossary is a familiar way for many humans to figure out what does a word in a particular context mean. We hypothesize that a system that can select a proper definition for a particular word occurrence can also naturally solve tasks related to word senses. To verify this hypothesis we developed a solution for the Multilingual and Cross-lingual Word-in-Context (MCL-WiC) task, that does not use any of the shared task data or other WiC data for training. Instead, it is trained to embed word definitions from English WordNet and word occurrences in English texts into the same vector space following an approach previously proposed by Blevins and Zettlemoyer (2020) for Word Sense Disambiguation (WSD). To estimate the similarity in meaning of two word occurrences, we compared different metrics in this shared vector space and found that L1-distance between normalized contextualized word embeddings outperforms traditionally employed cosine similarity and several other metrics. To solve the task for languages other than English, we rely on zero-shot cross-lingual transfer capabilities of the multilingual XLM-R masked language model. Despite not using MCL-WiC training data, in the shared task our approach achieves an accuracy of 89.5% on the English test set, which is only 4% less than the best system. In the multilingual subtask zero-shot cross-lingual transfer shows competitive results, that are within 2% from the best systems for Russian, French, and Arabic. In the cross-lingual subtask are within 2-4% from the best systems.

1 Introduction

SemEval-2021 Task 2 is a multilingual and cross-lingual word-in-context disambiguation task (MCL-WiC) for five different languages (Martelli et al., 2021). [1] Each example in the multilingual subtask consists of two sentences in English, Russian, French, Arabic, or Chinese language containing occurrences of the same target word. In the cross-lingual subtask each example consists of two sentences in different languages containing occurrences of two different target words. The participants were asked to detect whether those occurrences corresponded to the same or different meanings. These tasks are formalized as binary classification tasks. The datasets contain the same number of examples for each class. Accuracy is utilized as the main performance metric.

Recent SOTA approaches to the WiC task mainly include fine-tuning large universally pre-trained masked language models (Raffel et al., 2020; Liu et al., 2019) on labeled WiC datasets. Instead, we decided to train a system that selects the most appropriate definition for each word occurrence following an approach proposed by Blevins and Zettlemoyer (2020) for Word Sense Disambiguation (WSD) and experimented with different ways of adapting such system to the MCL-WiC task. During the evaluation period, we experimented with distances between probability distributions over word definitions but did not manage to achieve good results. But through the post-evaluation period, we switched to distances between the contextualized word embeddings of our gloss-informed language model and improved our results significantly achieving comparable results with the 2nd best system for French and 6th best system for Arabic in the multilingual subtask.

Our main interest was whether the word-in-context systems can benefit from using gloss information, provided for each possible sense of the word.

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

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 756–762
Bangkok, Thailand (online), August 5–6, 2021.

2 Background

Here we summarize prior work linking word occurrences and word definitions. One of the first approaches in this field (Lesk, 1986) calculated the lexical overlap between the context of a particular word occurrence and all possible definitions of this word. This approach did not take into account word synonymy or other lexical relations. The following work tried to combine state-of-the-art language models with glosses from some dictionaries.

One of such methods has been proposed by Kumar et al. (2019). Their EWISE system used a pre-training procedure for a gloss encoder, that learned knowledge graph embeddings from WordNet (Miller, 1995). After this pre-training, the authors froze the gloss encoder and started to train a context encoder with labeled WSD data. While the method of Kumar et al. (2019) requires relational information from a knowledge graph, the method proposed by Huang et al. (2019) relies fully on gloss information. The developed system jointly encodes the context with all possible glosses of the target word. The authors used a pre-trained BERT (Devlin et al., 2019) model as initialization for their encoder.

A similar approach has been proposed by Blevins and Zettlemoyer (2020), who trained two separate Transformer-based encoders for word occurrences (Context encoder) and word definitions (Gloss encoder), both initialized with BERT weights (Devlin et al., 2019). To represent a word occurrence, the outputs of the Context encoder for all of its subwords were averaged. To represent a definition, the output of the Gloss encoder from [CLS] token was taken. Finally, for a word occurrence and all of its definitions, the dot products between those outputs were calculated and the softmax function was applied to them, resulting in a probability distribution over possible word senses. The whole model was trained using cross-entropy loss to select the correct word sense on WSD data.

3 System overview

In order to learn sense-dependent representations of words, we pre-train our system on the Word Sense Disambiguation task. Following the BEM model (Blevins and Zettlemoyer, 2020), our system consists of two separate encoders: Context Encoder and Gloss Encoder.

Context encoder (T_c) takes a sentence $c = c_0, \ldots, c_{i-1}, w_c, c_{i+1}, \ldots, c_n$ containing a target word w_c to be disambiguated, where w_c is the i^{th} word in the sentence. The encoder then produces the target word representation:

$$r_{w_c} = T_c(c)[i]$$

For target words that are tokenized into multiple subword units, we average representations of these subwords.

Gloss encoder (T_g) takes as input a gloss g_s that defines a word sense s and encodes it as:

$$r_s = T_g(g_s)[0]$$

Taking the output from the first input token, which should be [CLS] for BERT or <s> for XLM-R.

We can score each of the possible senses $s \in S_w$, for a target word w_c by taking the dot product of r_{w_c} against every r_s for $s \in S_w$:

$$\phi(w_c, s) = r_{w_c}^T r_s$$

Both encoders were initialized with BERT or XLM-R weights. Then the whole system was pre-trained on English WSD data (Miller et al., 1994) with cross-entropy loss. We denote this pre-training procedure as Gloss Language Modeling (GLM) and compare it with pure Masked Language Modeling (MLM) pre-training. In both cases, the models were not fine-tuned on any MCL-WiC data.

3.1 Adaptation to the MCL-WiC task

As EWISE (Kumar et al., 2019) and BEM (Blevins and Zettlemoyer, 2020) systems work only with English data, we extend the proposed approach to the multilingual setting by replacing BERT with XLM-R model (Conneau et al., 2019). In the result section, we discuss how this affects the resulting performance.

Here we present two approaches to the final MCL-WiC task, one using distributions over possible word definitions and another exploiting different similarity measures between contextualized target word embeddings from the Context Encoder. The latter is also applicable to the contextualized word embeddings obtained from MLM pre-trained XLM-R, which we consider as a baseline.

3.1.1 Probability distribution over glosses

Here we exploit probability distributions $P(sense|w_{c_1})$ and $P(sense|w_{c_2})$ produced by

Sentence	gloss #1	gloss #2	gloss #3
dev.en-en.1, Meanings are the same			
No clause in a contract shall be interpreted as evading the responsibility of **superiors** under international law	one of greater rank or station or quality (0.82)	the head of a religious community (0.12)	a combatant who is able to defeat rivals (0.06)
In Senegal too, the customs officer and his **superiors** receive a premium in case of detecting and preventing smuggling	one of greater rank or station or quality (0.58)	the head of a religious community (0.3)	a combatant who is able to defeat rivals (0.09)
dev.en-en.16, Meanings are different			
During the fight both of them **tripped**, the author falling on the victim and stabbing him with the knife by accident	miss a step and fall or nearly fall (0.998)	cause to stumble (0.0009)	put in motion or move to act (0.0004)
The father of the child also cannot take the child to **trip** during the fostering duration, without permission of fosterer	make a trip for pleasure (0.9)	miss a step and fall or nearly fall (0.06)	get high, stoned, or drugged (0.02)

Table 1: Examples from the MCL-WiC development set with 3 most probable glosses of the target word predicted by our system. Word in **bold** is the target word. The rounded probabilities for each of the meanings are given in parentheses.

our WSD system for words w_{c_1} and w_{c_2} in their contexts c_1 and c_2. w_{c_1} and w_{c_2} have the same lemma and consequently have the same set of possible meanings S_w from the vocabulary.

Gloss match prob: The probability that two word occurrences, w_{c_1} and w_{c_2}, have the same meaning (positive class) is calculated as:

$$P(1|w_{c_1}, w_{c_2}) = \sum_{s_i \in S_w} P(s_i|w_{c_1}) \cdot P(s_i|w_{c_2})$$

Gloss JSD: As an alternative measure of word similarity in context, we compute Jensen–Shannon divergence between two distributions $P(sense|w_{c_1})$ and $P(sense|w_{c_2})$.

Because these methods rely on gloss information, we need a vocabulary to find them. As for English, we can easily use WordNet (Miller, 1995), it becomes problematic to find definitions for other languages. To counteract this obstacle during the competition we used the following procedure. First, we translated all samples from other languages to English via machine translation [2]. Then we generated all possible translations of the target word with Yandex Translation API [3] and word2word library [4]. Finally, we tried to find one of the possible target word translations in the translated sentence. If there was no match for both sentences, we predicted True, if a match was only for one, we predicted False. In case of more than one match in the translated sentence, we took the first one. In the end, we had two occurrences of possibly different translations of the target word into English and could use previous metrics. But unlike the original English method where we need to disambiguate between meanings of the same word, here we build the distribution over all possible glosses of all possible translations of the target word.

3.1.2 Similarity between contextualized word embeddings

In this subsection, we propose methods that fully rely on the Context encoder and thus do not require any additional vocabulary or glosses. We achieve such generalization by using only outputs from the trained Context encoder.

Cosine: Cosine similarity between outputs of the encoder.

Euclidian+norm: Euclidian distance between L2 normalized outputs of the encoder.

Manhattan+norm: Manhattan distance between L1 normalized outputs of the encoder.

3.2 Threshold selection

As MCL-WiC is a binary classification problem, we need to transform our continuous similarity measures into binary predictions. We select the best threshold with grid search on one of the following datasets.

1. **(xx dev)** The threshold is selected on the development set of the target language.

2. **(en dev)** The threshold for all languages is selected on the English development set.

[2] https://huggingface.co/transformers/model_doc/marian.html

[3] https://yandex.com/dev/dictionary/

[4] https://github.com/kakaobrain/word2word

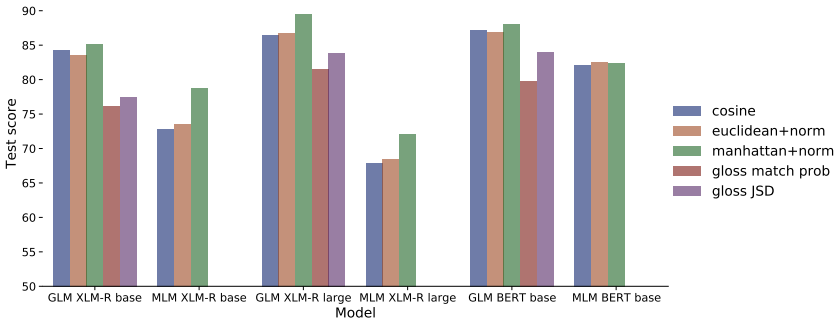

Figure 1: Comparison of pre-training methods and similarity measures on the English MCL-WiC test set.

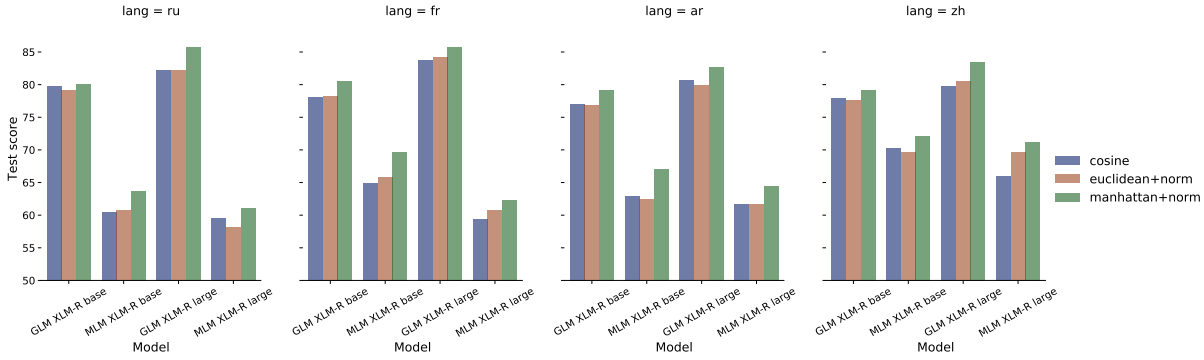

Figure 2: Comparison of pre-training methods and similarity measures on MCL-WiC test sets.

3. **(semcor)** Instead of utilizing MCL-WiC development data, we tried estimating the optimal threshold on the training SemCor dataset. Based on sense annotations, we constructed 30K sentence pairs in WiC format. This resulted in a nearly-balanced WiC dataset, which was used for grid search.

4. **(cl trials)** The threshold for the cross-lingual subtask is selected on the concatenated cross-lingual MCL-WiC trial sets.

4 Experimental setup

Besides choosing the best final predictor for MCL-WiC, we also experimented on encoder initialization. In the result section, we compare the performance of the models, initialized with BERT base (Devlin et al., 2019), XLM-R base, and XLM-R large (Conneau et al., 2019). We trained our models on the English SemCor dataset (Miller et al., 1994) with glosses from the WordNet 3.0 (Miller, 1995). Systems based on the XLM-R base and XLM-R large (Conneau et al., 2019) were trained 20 and 10 epochs respectively. Following standard practices, we used SemEval-2007 (Pradhan et al., 2007) as our development set for early stopping. For the system with BERT-base, we used the originally provided checkpoint by Blevins and Zettlemoyer (2020).

5 Results

5.1 Similarity measures

The results of the experiments with different similarity measures for English and non-English languages are given in Figure 1 and Figure 2 respectively. Figure 1 shows that approaches based on GLM context outputs strongly outperform methods based on distributions over senses from the vocabulary.

Figures 1, 2 also provide an observation, that almost for any language and model Manhattan+norm distance shows the best results. The only exception is the MLM BERT base model, where Euclidian+norm performs slightly better.

5.2 GLM vs MLM

Figure 4 shows the gap between GLM and pure MLM pre-training. Experiments show that models trained with GLM procedure strongly outperform their MLM counterparts in every language and with any base model.

Model	en	ru	fr	ar	zh
our post-evaluation results					
GLM XLM-R base - Manhattan+norm	85.1	80.1	80.5	79.2	79.1
MLM XLM-R base - Manhattan+norm	78.8	63.6	69.6	67.1	72.1
GLM XLM-R large - Manhattan+norm (en dev)	**89.5**	**85.7**	**86.5**	**84.2**	**83.5**
GLM XLM-R large - Manhattan+norm (semcor)	87.8	82.7	84.2	80.9	80.8
GLM XLM-R large - Manhattan+norm	**89.5**	**85.7**	85.7	82.6	**83.5**
MLM XLM-R large - Manhattan+norm	72.1	61.1	62.3	64.5	71.1
GLM BERT base - Manhattan+norm	88	-	-	-	-
MLM BERT base - Euclidean+norm	82.5	-	-	-	-
our submissions					
GLM BERT base - Gloss JSD	86.4	-	-	-	-
GLM BERT base (MT) - Gloss JSD	86.4	68.3	69.2	50.1	64.1
best submissions					
Best for each lang.	**93.3**	**87.4**	**87.5**	**84.8**	**91**

Table 2: Best test score for each of the proposed systems. *(MT)* states for the Machine Translation technique, described in Section 3.1. The threshold for binary classification was calculated either on the English dev set *(en dev)*, or on a part of SemCor dataset *(semcor)*, or on the dev set, corresponding to the target language (all others). *Best for each lang.* stands for the best results of the competition for each of the languages.

Surprisingly, the XLM-R base model outperforms the large one when pre-trained with MLM objective only. However, after GLM pre-training the large model performs significantly better than the base model. We suspect that this is due to the strong grammatical bias of contextualized word embeddings after MLM pre-training also observed by Laicher et al. (2021). It is easier to correctly predict the grammatical form of a masked word in a particular context than the exact lemma of that word. Thus, the model is much more confident in the grammatical form resulting in distant embeddings for the same word in the same sense but in different grammatical forms. Since the large model shall better optimize the MLM objective, its embeddings likely contain stronger grammatical component hiding the sense component relevant for the MCL-WiC task. Since word definitions correspond to word senses and not word forms, the GLM objective helps eliminating the irrelevant grammatical component from contextualized embeddings.

5.3 Correlation between WSD and MCL-WiC performance

We also suspected that during pre-training on the English WSD data, GLM models can overfit to English texts and partially lose their cross-lingual transferability. Thus, the best epoch checkpoint based on WSD development set may not be the best in terms of MCL-WiC. In Figure 3 we show how MCL-WiC accuracy for each language and WSD F1 score for English change during training. Multilingual WiC performance for epoch 0 stands for the MLM model without any GLM training.

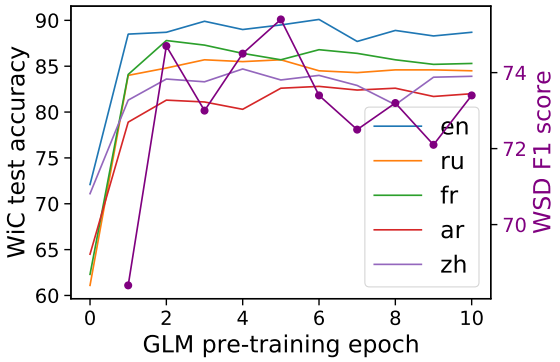

Figure 3: Test MCL-WiC score and SemEval07 dev F1 score for the XLM-R large model depending on the epoch.

The results show that choosing the best checkpoint by WSD F1 score, we get nearly optimal results on MCL-WiC for each language, except for French.

5.4 Interpreting system predictions

As our system embeds word definitions from WordNet (Miller, 1995) and word occurrences into the same vector space, we can search for the nearest definitions for each occurrence of the target word. In Table 1 we show some examples from the development set with top3 senses (glosses) predicted by our system for each occurrence. We used GLM XLM-R large as a backbone for this purpose.

5.5 Overall multilingual results

Table 2 shows overall results on the competition's test set. The submission *GLM BERT base - Gloss JSD* sent during the competition employed English BERT base (Devlin et al., 2019) backbone in the

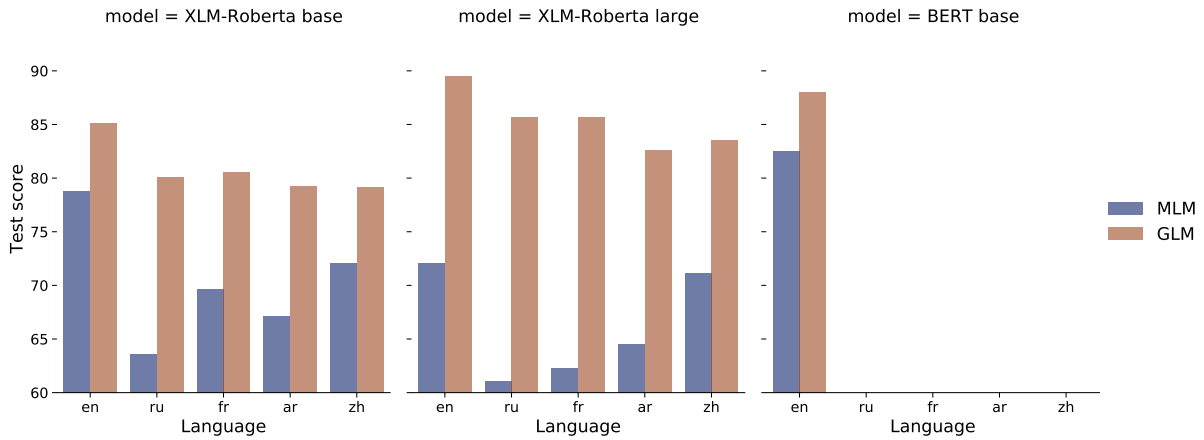

Figure 4: Test score for the best GLM and MLM models for each language.

WSD model, which was applicable to English data only. The submission *GLM BERT base (MT) - Gloss JSD* exploits the same basic idea but extends the first approach to other languages with the translation technique, described in Section 3.1. For both submissions, due to a mistake, we used the GLM model trained for only one epoch on SemCor (Miller et al., 1994).

As we can see from the table 2, our best system on every language is GLM pre-trained XLM-R large model with Manhattan+norm distance and the threshold selected on the English MCL-WiC development set. This system shows rather strong results achieving the performance of the 2nd best system for French and the 6th best system for Arabic. Since this system does not use any non-English resources for training, we suspect that it will work for a variety of other languages on which XLM-R was initially pre-trained, though the performance may vary.

5.6 Cross-lingual results

Table 3 shows our post-evaluation results on the cross-lingual subtask of MCL-WiC. The threshold selected on the English dev set does not transfer to the cross-lingual test sets, unlike multilingual test sets. This is due to larger distances between contextualized embeddings returned by XLM-R for word occurrences in different languages. Selecting threshold on the concatenation of cross-lingual trial sets works much better. However, since there are only 32 cross-lingual examples, there is a wide interval of optimal thresholds giving the same accuracy on trial. Our implementation selected the smallest one, however, some larger thresholds resulted in a significant decrease in performance.

Model	en-ar	en-fr	en-ru	en-zh
our post-evaluation results				
(en dev)	77.6	81.5	81.8	78.9
(cl trials)	**85.2**	**85.5**	**87.2**	**89.2**
best submissions				
Best for each pair	**89.1**	**89.1**	**89.4**	**91.2**

Table 3: Post-evaluation test scores for the cross-lingual subtask. For each of our systems, we used GLM XLM-R large model and Manhattan+norm distance. *Best for each pair* stands for the best results of the competition for each pair of languages individually.

Thus, a larger cross-lingual development set is required for robust selection of the threshold.

6 Conclusion

In this paper, we presented Gloss Language Modeling (GLM) procedure as a pre-training strategy for MCL-WiC systems. We have shown that this procedure improves multilingual WiC performance on all languages for both XLM-R and BERT backbones.

Apart from that, we proposed an interpretable zero-shot multilingual WiC algorithm which does not require any labeled data for the multilingual WiC task except for the threshold selection, which can be performed using only English development data without loss of accuracy for other languages. We also found that L1-distance between normalized contextualized word embeddings outperforms traditionally employed cosine distance.

Acknowledgments

This research was supported through computational resources of HPC facilities at NRU HSE.

References

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 Association for Computational Linguistics*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. 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 3507–3512, Hong Kong, China. Association for Computational Linguistics.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681, Florence, Italy. Association for Computational Linguistics.

Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and improving BERT performance on lexical semantic change detection.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, page 24–26, New York, NY, USA. Association for Computing Machinery.

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.

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 Workshop on Semantic Evaluation (SemEval-2021)*.

George A. Miller. 1995. WordNet: A lexical database for English. *Commun. ACM*, 38(11):39–41.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, page 240–243, USA. Association for Computational Linguistics.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.