SkoltechNLP at SemEval-2021 Task 2: Generating Cross-Lingual Training Data for the Word-in-Context Task

Anton Razzhigaev[1], **Nikolay Arefyev**[2,3,4], **and Alexander Panchenko**[1]
[1]Skolkovo Institute of Science and Technology, Russia
[2]Samsung Research Center Russia, Russia
[3]Lomonosov Moscow State University, Russia
[4]HSE University, Russia
{anton.razzhigaev,a.panchenko}@skoltech.ru
narefjev@cs.msu.ru

Abstract

In this paper, we present a system for the solution of the cross-lingual and multilingual word-in-context disambiguation task. Task organizers provided monolingual data in several languages, but no cross-lingual training data were available. To address the lack of the officially provided cross-lingual training data, we decided to generate such data ourselves. We describe a simple yet effective approach based on machine translation and back translation of the lexical units to the original language used in the context of this shared task. In our experiments, we used a neural system based on the XLM-R (Conneau et al., 2020), a pre-trained transformer-based masked language model, as a baseline. We show the effectiveness of the proposed approach as it allows to substantially improve the performance of this strong neural baseline model. In addition, in this study, we present multiple types of the XLM-R based classifier, experimenting with various ways of mixing information from the first and second occurrences of the target word in two samples.

1 Introduction

The goal of the second task of SemEval-2021 (Martelli et al., 2021) is to perform multilingual and cross-lingual word-in-context disambiguation. More specifically, participants are asked to distinguish whether the meanings of a target word in two provided contexts are the same or not. Organizers provided a training set of 8 000 English language (en-en) context pairs and validation sets of 1 000 context pairs for English-English (en-en), French-French (fr-fr), Russian-Russian (ru-ru), Arabic-Arabic (ar-ar), and Chinese-Chinese (zh-zh) languages. Since no cross-lingual training data were provided, except for a very small trial set barely usable for training, we decided to venture into generating such data automatically.

Essentially, the given task is a binary classification problem. The first question was which supervised model to use for the classification of context pairs. Recently, pre-trained masked language models such as BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) have been used to reach promising results in a variety of similar NLU classification tasks. Thus, we decided to make use of contextualized embeddings from XLM-R, which provides multilingual-lingual embeddings for more than 100 languages, covering all language pairs of interest in the shared task. In all our experiments, this model is used as the backbone.

A straightforward way of solving tasks where two contexts are to be compared, as the word-in-context tasks, is to use deep contextualized embeddings and train a classifier over these embeddings as has been explored in the original monolingual word-in-context task (Pilehvar and Camacho-Collados, 2019). Note that commonly embeddings of two contexts are simply concatenated (Ma et al., 2019) and this operation is asymmetric. In our work, we explored various symmetric ways of aggregating embeddings from two contexts.

The contributions of our work are two-fold. First, we present a simple yet effective method for the generation of cross-lingual training data, showing that it can substantially improve the performance compared to the model trained using monolingual data. Second, we test various ways of encoding two input target word occurrences contexts using the XLM-R model.

2 Baseline Supervised WiC System

Massively multilingual transformers pretrained with language modeling objectives XLM-R were shown to be useful for zero-shot cross-lingual transfer in NLP (Lauscher et al., 2020). As a baseline, we rely on a supervised system that takes as an

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 157–162
Bangkok, Thailand (online), August 5–6, 2021.

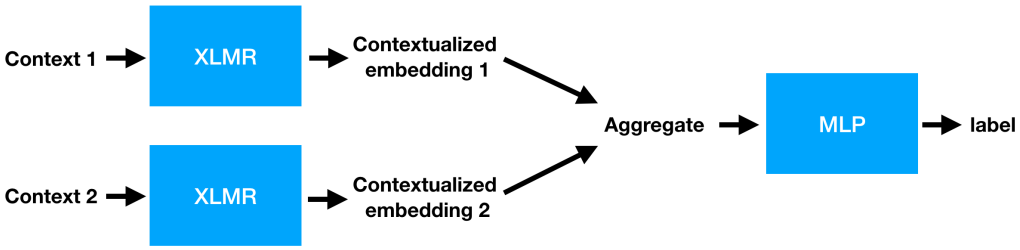

Figure 1: Principal scheme of the supervised model used in our experiments. Context pairs are sent to XLM-R base and then contextualized embeddings of target words are merged and sent to MLP which outputs the prediction probability of them having the same meaning. XLM-R is frozen.

input two sentences and spans corresponding to occurrences of the target word. Pre-trained multilingual encoders are used to represent sentences in different languages in the same space.

2.1 Multilingual Sentence Encoder

We use XLM-R masked language model (Conneau et al., 2020) as the basis in our experiments as it supports all required languages by the shared task. This is a multilingual transformer-based masked language model pre-trained on a large corpus consisting of texts from the Web in 100 languages. This model is a strong baseline on various NLU tasks. Besides, our preliminary experiments have shown that it is capable of encoding sentences written in different languages in the same vector space. This property, therefore, is crucial as it allows similar methods, which were used to successfully solve the monolingual word-in-context task in the past (Pilehvar and Camacho-Collados, 2019).

Figure 1 presents the overall schema of the model used in our experiments. The XLM-R model is used for obtaining contextualized embeddings of the target words, while a multi-layered perceptron is used to perform the classification. We thoroughly tested various meta-parameters of this architecture. Different aggregation methods are presented in the following section.

2.2 Symmetric Aggregation of Deep Contextualized Embeddings

Each training example consists of two contexts with marked target words and a label representing these words being in the same or different meanings. In our approach, both contexts are sent to XLM-R, and then contextualized embeddings for target words (averaged activations from two last layers) are extracted and merged into one embedding with the following symmetric procedure: concatenate element-wise product of two embeddings and the absolute value of the element-wise difference of two embeddings. This helps to obtain a vector containing deep contextualized representation of a target word in both contexts. Then this merged embedding is sent to a 3-layer MLP which outputs the probability of two words been in the same senses (Figure 1).

More specifically, we test different ways of aggregating embeddings from two contexts. We conducted several experiments, including two asymmetric aggregation approaches and four symmetric. Let $\vec{a} = \{a_1, ...a_n\}$ be the contextualized embedding of a target word from the first context and $\vec{b} = \{b_1, ...b_n\}$ – from the second.

The tested two following commonly used asymmetric approaches of merging two embeddings:

1. Concatenating of embeddings:
   $\vec{c} = \{a_1, ..., a_n, b_1, ..., b_n\}$
2. Difference of embeddings:
   $\vec{c} = \{a_1 - b_1, ..., a_n - b_n\}$

Besides, we tested four symmetric approaches to embedding aggregation listed below:

1. Sum of embeddings:
   $\vec{c} = \{a_1 + b_1, ..., a_n + b_n\}$
2. Elementwise product of embeddings:
   $\vec{c} = \{a_1 \cdot b_1, ..., a_n \cdot b_n\}$
3. Absolute value of difference of embeddings:
   $\vec{c} = \{|a_1 - b_1|, ..., |a_n - b_n|\}$
4. Concatenation of variants 2 and 3:
   $\vec{c} = \{a_1 \cdot b_1, ..., a_n \cdot b_n, |a_1 - b_1|, ..., |a_n - b_n|\}$

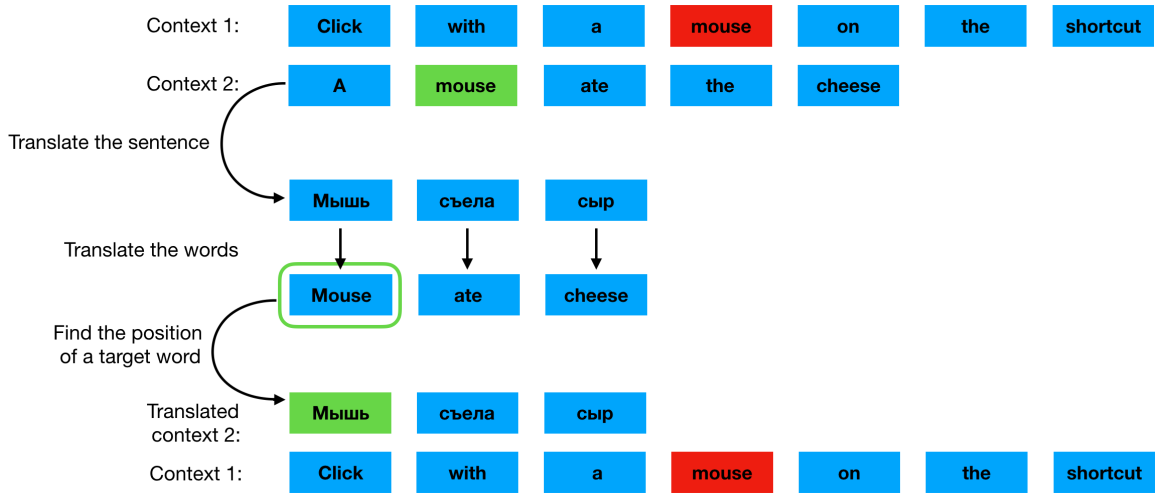

Figure 2: An illustration of the cross-lingual data generation. Given two sentences, we pick one and translate it to a target language. Then in order to find the position of a target word, every lexical unit is back-translated and compared with a target lemma. If the target word is found, the translated sentence is used in addition to the second sentence from the initial pair as a new cross-lingual training example.

3 Generation of Cross-lingual Training Data using Machine Translation

In this section, we describe a machine-translation-based method for the generation of synthetic training data for the cross-lingual word-in-context task to address the lack of cross-lingual training data usable for the supervised model described above.

3.1 Method

We suggest the forward-backward translation approach, which helps not just to translate a sentence but to identify the position of a target word which is essential for the word-in-context task.

We decided to use the provided 8 000 English-English pairs of texts and translate them to the desired languages. But there is a difficulty: after translation the position of target word in the context is unknown, or even target word is replaced by several words like in the following example of Russian-English translation (the target words are underlined):

- ru: “налей кипяток в стакан”
- en: “pour boiling water into a glass”

In our experiments, we filter similar examples which do not have a uniword translation of a target word.

Overall, our algorithm amounts to the following procedure:

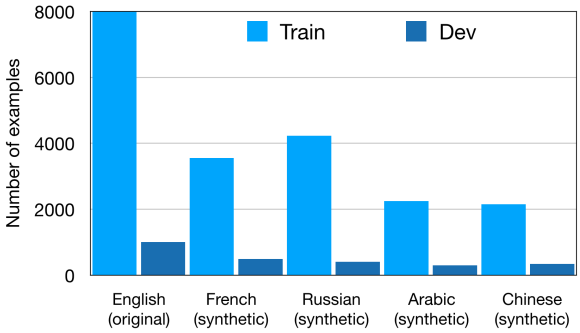

Figure 3: Amount of the English training/development data and amount of synthetic cross-lingual data generated from it.

1. Translate a sentence from the source language to a target with a neural machine translation.[1]
2. Back translate every word independently without a context. For the translation of single words, we use the word2word[2] library
3. If there is a target lemma in the list of back-translated words, then the lemma index in the back-translated words list is the index of the target word in the translated sentence.
4. If there is no target lemma in the list of back-translated words, then we do not use this sentence.

[1] https://github.com/ssut/py-googletrans (Google translate Python API)
[2] https://github.com/kakaobrain/word2word

label: False
text1: Her parents allegedly ***received*** threats from the police to make them drop the charges against the soldiers.
text2: Специальный докладчик ***получил*** информацию о положении в тюрьмах из государственных и негосударственных источников.

label: False
text1: Similarly, in the ***event*** of a transplant, the deceased donor must continue to be entitled to respect for his body.
text2: Следует прилагать усилия для разработки программ обмена и даже спортивных ***мероприятий***, таких как футбольные матчи.

label: True
text1: The Committee recommends such a ***procedure*** be reflected in the provisions contained in the proposed draft protocol.
text2: Его ***процедуры*** незамедлительных действий также заслуживают высокой оценки.

label: True
text1: During a summer the two students started doing their own research on a ***roulette*** wheel which they had bought.
text2: Используя MACS для записи вращений стола ***рулетки*** с течением времени, компьютер может предсказывать будущие результаты.

Figure 4: Examples of generated synthetic cross-lingual data.

Method	Type	fr-fr
Concatenating	A	67.0 ± 2.0
Difference	A	65.4 ± 1.3
Summation	S	79.6 ± 1.3
Elementwise product	S	81.8 ± 1.4
Absolute difference	S	81.5 ± 1.8
Concat of symmetric	S	$\mathbf{82.1} \pm 1.4$

Table 1: Symmetric (S) vs asymmetric (A) ways of merging XLMR-large contextualized embeddings. Concatenation of symmetric: concatenation of elementwise multiplication and absolute difference.

A schematic illustration of our algorithm is presented in Figure 2.

3.2 Generation Result

The synthetic examples for English-Russian are presented in Figure 4. The first two sentence pairs (with the False, F label) represent negative training examples, i.e., pairs of sentences in which target words are used in different senses (across languages). The last two sentence pairs (with the True, T label) represent contexts where words are used in the same sense. As one may observe, the generated examples are semantically coherent, and the position of the target word was identified correctly using our back-translation heuristic.

The overall amount of generated training cross-lingual examples for each language compared to the amount of initial English language data presented in Figure 3. The unequal number of samples is due to the translation errors and the fact that back translation does not always point to the original word. That is why we also present results for the fixed sizes of synthetic datasets for each language in the Table 2.

4 Experiments and Results

Below we report the results of the two setups of this shared task: multi- and cross-lingual settings. We train the model six times; reporting mean and standard deviation of accuracy on the test dataset.

4.1 Results on Various Embedding Aggregation Methods

All embedding aggregation methods were tested on the French language development set, been trained on the English training set. The experimental results are presented in Table 1. Experimental results demonstrate that the suggested symmetric aggregation of embeddings is a better choice for such symmetric problems like two context comparisons than a common asymmetric aggregation. We suppose that this experimental fact is caused by the symmetric nature of a comparison problem and hence all similar tasks should exploit symmetrically merged embeddings.

4.2 Results on Multilingual Datasets

In a multilingual setting, context pairs are provided in four languages, but pairs are written in the same language. As XLM-R provides contextualized text representations in the same space for different languages, we supposed that our XLM-R based model should work in a zero-shot setting: being trained on only one language shows decent results on other languages. To verify our hypothesis, we conducted the following experiments:

1. Training only on 8 000 MCL-WiC English context pairs (zero-short setting).

2. Training on 8 000 MCL-WiC English context pairs (from the training set) + 5 000 multilanguage pairs (from development set).

The results are presented in Table 2: substantially higher results than the random baseline (50

Training set	en-en	fr-fr	ru-ru	ar-ar	zh-zh
English train data	87.5 ± 0.9	82.1 ± 1.4	78.9 ± 1.7	69.2 ± 1.9	65.2 ± 1.5
English train and multilingual dev	$\mathbf{89.9} \pm 0.8$	$\mathbf{84.1} \pm 1.5$	$\mathbf{86.5} \pm 1.1$	$\mathbf{72.4} \pm 1.3$	$\mathbf{70.2} \pm 1.0$

Table 2: Results on test data in multi-lingual setting.

Training set	en-fr	en-ru	en-ar	en-zh
English train data	64.1 ± 2.7	61.4 ± 2.1	59.1 ± 2.1	52.9 ± 1.3
English train and multilingual dev	66.5 ± 1.2	62.0 ± 1.0	58.9 ± 1.8	52.1 ± 0.7
Synthetic for each language (fixed)	70.1 ± 2.1	69.7 ± 1.9	$\mathbf{63.1} \pm 2.1$	$\mathbf{60.1} \pm 1.4$
Synthetic for each language (full)	72.6 ± 2.0	71.4 ± 1.4	62.7 ± 2.1	$\mathbf{60.1} \pm 1.4$
All data	$\mathbf{73.5} \pm 1.9$	$\mathbf{72.8} \pm 1.5$	58.8 ± 1.7	52.1 ± 0.6

Table 3: Results on test data in cross-lingual setting.

percent) are obtained. Note that in this case, the dataset is balanced, so the most frequent class classifier is equivalent to the random one. This confirms the fact that a zero-shot transfer using XLM-R is possible.

4.3 Results on Cross-lingual Datasets

In a cross-lingual setting, context pairs are provided in four languages, and pairs are written in different languages. The main challenge of the task is cross-lingual Word-in-Context disambiguation. We approach this task from two sides: zero-shot learning capabilities of multilingual XLM-R based systems and generation with a machine translation of cross-lingual synthetic training data. To verify that zero-shot learning works in a cross-lingual setting and synthetically generated data improves the results in cross-lingual tests, we performed the following experiments:

1. Training only on 8 000 MCL-WiC English context pairs (zero-short setting).
2. Training on 8 000 MCL-WiC English context pairs + 10 000 multi-language pairs.
3. Training on synthetic cross-lingual examples. Training and testing each language separately.
4. Training on all data including MCL-WiC train, development sets, and synthetic cross-lingual data for all languages simultaneously.

Results are presented in the Table 3. The best results for Russian and French are obtained using all the available data, including the generated synthetic dataset. For Arabic and Chinese, the best results are obtained using synthetic data only. Overall, performance in all settings for Chinese and Arabic is substantially lower. This may be due to the more complex morphological structure of these languages and the way how the XLM-R pre-trained model handles it (while the European languages like French and Russian have similar alphabet structures). Overall, the experiments suggest the usefulness of the generated synthetic data for the solution of the cross-lingual word-in-context task.

5 Conclusion

In this paper, we presented a solution to the cross-lingual word-in-context task. The main challenge of this task, as formulated by the organizers, is the lack of explicit training data. To address it, we developed a way of generating synthetic cross-lingual data for the word-in-context disambiguating task; we demonstrate the positive influence of such synthetic data on the performance of a model on test datasets.

As the baseline model in our experiments, a supervised model based on XLM-R pre-trained language model (Conneau et al., 2020) was used. We performed tests of various settings based on this model and demonstrated that symmetric aggregation of embeddings for context comparison tasks outperforms asymmetric ways on zero-shot and supervised settings.

The code and the produced data, enabling reproducing our experiment, are available online.[3]

[3] https://github.com/skoltech-nlp/cross-lingual-wic

References

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.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Xiaofei Ma, Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. 2019. Universal text representation from BERT: an empirical study. *CoRR*, abs/1910.07973.

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)*.

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.