# HIT-SCIR at MMNLU-22: Consistency Regularization for Multilingual Spoken Language Understanding

**Bo Zheng, Zhouyang Li, Fuxuan Wei, Qiguang Chen, Libo Qin, Wanxiang Che***
Harbin Institute of Technology
{bzheng,zhouyangli,fxwei,qgchen,lbqin,car}@ir.hit.edu.cn

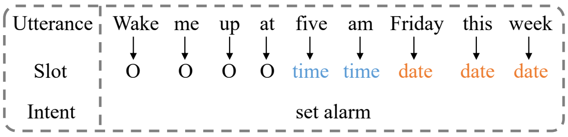
## Abstract

Multilingual spoken language understanding (SLU) consists of two sub-tasks, namely intent detection and slot filling. To improve the performance of these two sub-tasks, we propose to use consistency regularization based on a hybrid data augmentation strategy. The consistency regularization enforces the predicted distributions for an example and its semantically equivalent augmentation to be consistent. We conduct experiments on the MASSIVE dataset under both full-dataset and zero-shot settings. Experimental results demonstrate that our proposed method improves the performance on both intent detection and slot filling tasks. Our system[1] ranked 1st in the MMNLU-22 competition under the full-dataset setting.

## 1 Introduction

The MMNLU-22 evaluation focuses on the problem of multilingual natural language understanding. It is based on the MASSIVE dataset (FitzGerald et al., 2022), a multilingual spoken language understanding (SLU) dataset with two sub-tasks, including *intent detection* and *slot filling*. Specifically, given a virtual assistant utterance in an arbitrary language, the model is designed to predict the corresponding intent label and extract the slot results. An English example is illustrated in Figure 1.

Fine-tuning pre-trained cross-lingual language models allows task-specific supervision to be shared and transferred across languages (Conneau and Lample, 2019; Conneau et al., 2020; Xue et al., 2021). This motivates the two setting for the MMNLU-22 evaluation, namely the *full-dataset* setting and the *zero-shot* setting. Participants are allowed to use training data in all languages under the full-dataset setting, while they can only access the English training data under the zero-shot setting.



| Utterance | Wake | me | up | at | five | am | Friday | this | week |
|---|---|---|---|---|---|---|---|---|---|
| Slot | O | O | O | O | time | time | date | date | date |
| Intent | set alarm | | | | | | | | |

Figure 1: An English example from the MASSIVE dataset. The slot label 'O' stands for the 'Other' label.

The latter is also called zero-shot cross-lingual SLU in previous work (Qin et al., 2020, 2022).

Cross-lingual data augmentation methods have been proven effective to improve cross-lingual transferability, e.g., code-switch substitution (Qin et al., 2020) and machine translation (Conneau and Lample, 2019; Singh et al., 2019). Most previous work directly utilizes the data augmentations as additional training data for fine-tuning. However, they ignore the inherent correlation between the original example and its semantically equivalent augmentation, which can be fully exploited with the *consistency regularization* (Zheng et al., 2021b). The consistency regularization enforces the model predictions to be more consistent for semantic-preserving augmentations.

Motivated by this, we propose to apply consistency regularization based on a hybrid data augmentation strategy, including data augmentation of machine translation and subword sampling (Kudo, 2018). We use machine translation augmentation to align the model predictions of the intent detection task. Meanwhile, subword sampling augmentation is used to align the model predictions of both intent detection and slot filling tasks. The proposed method consistently improves the SLU performance on the MASSIVE dataset under both full-dataset and zero-shot settings. It is worth mentioning that our system ranked 1st in the MMNLU-22 competition under the full-dataset setting. We achieved an exact match accuracy of 49.65 points, outperforming the 2nd system by 1.02 points.

*Email corresponding.

[1] The code will be available at https://github.com/bozheng-hit/MMNLU-22-HIT-SCIR.

*Proceedings of the Massively Multilingual Natural Language Understanding Workshop (MMNLU-22)*, pages 35 - 41
December 7, 2022 ©2022 Association for Computational Linguistics

# 2 Background

## 2.1 Task Description

The task of SLU is that given an utterance with a word sequence $\boldsymbol{x} = (x_1, ..., x_n)$ with length $n$. The model is required to solve two sub-tasks. The intent detection task can be seen as an utterance classification task to decide the intent label $o^I$, and the slot filling task is a sequence labeling task that generates a slot label for each word in the utterance to obtain the slot sequence $\boldsymbol{o}^S = (o^S_1, ..., o^S_n)$.

## 2.2 Dataset Description

The MASSIVE dataset is composed of realistic, human-created virtual assistant utterance text spanning 51 languages, 60 intents, 55 slot types, and 18 domains (FitzGerald et al., 2022). There are 11,514 training utterances for each language. For the full-dataset setting, all training data can be used. For the zero-shot setting, only English training data can be used, yet we can translate them into other languages using commercial translators. There are 2,033, 2,974, and 3,000 utterances for each language in the development, test, and evaluation set, respectively. The average performance in all languages should be reported under the full-dataset setting. Meanwhile, the average performance in all languages except English should be reported under the zero-shot setting.

## 2.3 Related Work

Pre-trained cross-lingual language models (Conneau and Lample, 2019; Conneau et al., 2020; Chi et al., 2021a,b, 2022; Xue et al., 2021) encode different languages into universal representations and significantly improve cross-lingual transferability. These models usually consist of a multilingual vocabulary (Conneau and Lample, 2019; Conneau et al., 2020; Xue et al., 2021; Zheng et al., 2021a) and a Transformer model (Vaswani et al., 2017).

A simple yet effective way to improve cross-lingual fine-tuning is to populate the training data with cross-lingual data augmentation (Conneau et al., 2020). Singh et al. (2019) replace a segment of source language input text with its translation in another language as data augmentation. Qin et al. (2020) randomly replace words in the source-language training example with target-language words using the bilingual dictionaries. Then the model is fine-tuned on the generated code-switched data. Instead of directly treating cross-lingual data augmentation as extra training data, Zheng et al. (2021b) proposed to better use data augmentations based on consistency regularization.

# 3 Method

Given the input utterance $\boldsymbol{x} = (x_1, ..., x_n)$ with length $n$ and the corresponding intent label $o^I$ and slot labels $\boldsymbol{o}^S = (o^S_1, ..., ...o^S_n)$ from training corpus $\mathcal{D}$, we define the loss for the two sub-tasks of SLU in our fine-tuning process as:

$$\mathcal{L}_I = \sum_{(\boldsymbol{x}, o^I) \in \mathcal{D}} \mathrm{CE}(f_I(\boldsymbol{x}), o^I),$$

$$\mathcal{L}_S = \sum_{(\boldsymbol{x}, \boldsymbol{o}^S) \in \mathcal{D}} \mathrm{CE}(f_S(\boldsymbol{x}), \boldsymbol{o}^S),$$

where $\mathcal{L}_I$ and $\mathcal{L}_S$ stand for the intent detection task and the slot filling task, $f_I(\cdot)$ and $f_S(\cdot)$ denote the model which predicts task-specific probability distributions for the input example $\boldsymbol{x}$, CE$(\cdot, \cdot)$ denotes cross-entropy loss.

## 3.1 Consistency Regularization

In order to make better use of data augmentations, we introduce the consistency regularization used in Zheng et al. (2021b), which encourages consistent predictions for an example and its semantically equivalent augmentation. We apply consistency regularization on intent detection and slot filling tasks, which is defined as follows:

$$\mathcal{R}_I = \sum_{\boldsymbol{x} \in \mathcal{D}} \mathrm{KL}(f_I(\boldsymbol{x}) \| f_I(\mathcal{A}(\boldsymbol{x}, z))),$$

$$\mathcal{R}_S = \sum_{\boldsymbol{x} \in \mathcal{D}} \mathrm{KL}(f_S(\boldsymbol{x}) \| f_S(\mathcal{A}(\boldsymbol{x}, z))),$$

$$\begin{aligned} \mathrm{KL}_\mathrm{S}(P \| Q) = \ & \mathrm{KL}(\mathrm{stopgrad}(P) \| Q) + \\ & \mathrm{KL}(\mathrm{stopgrad}(Q) \| P) \end{aligned}$$

where $\mathrm{KL}_\mathrm{S}(\cdot\|\cdot)$ is the symmertrical Kullback-Leibler divergence, $\mathcal{A}(\boldsymbol{x}, z)$ denotes the augmented version of input utterance $\boldsymbol{x}$ with data augmentation strategy $z$. The regularizer encourages the predicted distributions of the original training example and its augmented version to agree with each other. The $\mathrm{stopgrad}(\cdot)$ operation[2] is used to stop back-propagating gradients, which is also employed in (Jiang et al., 2020; Liu et al., 2020; Zheng et al., 2021b).

## 3.2 Data Augmentations

We consider two types of data augmentation strategies for our consistency regularization method, including subword sampling and machine translation.

[2] Implemented by `.detach()` in PyTorch.

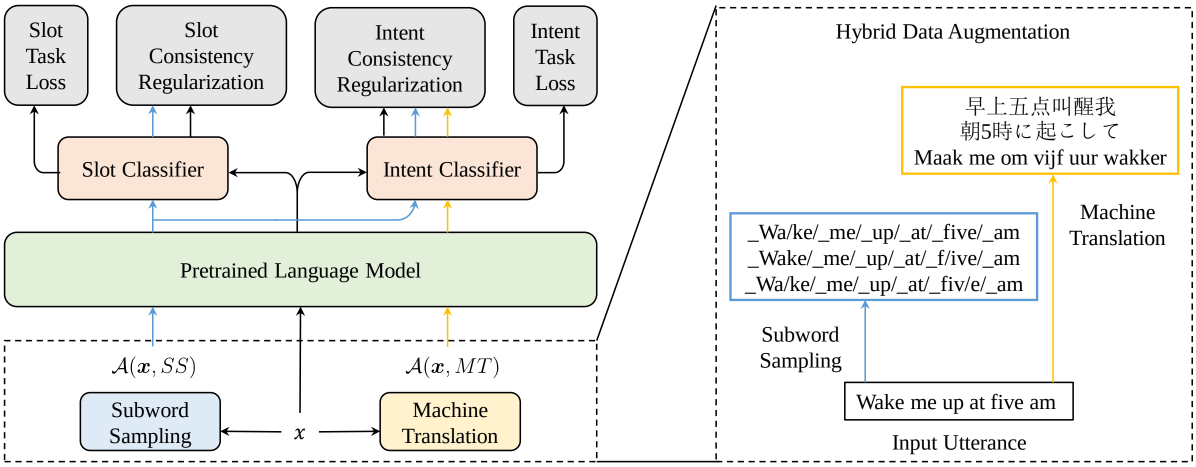

Slot Task Loss
Slot Consistency Regularization
Intent Consistency Regularization
Intent Task Loss
Slot Classifier
Intent Classifier
Pretrained Language Model
$\mathcal{A}(\boldsymbol{x}, SS)$
$\mathcal{A}(\boldsymbol{x}, MT)$
Subword Sampling
$x$
Machine Translation
Hybrid Data Augmentation
早上五点叫醒我
朝5時に起こして
Maak me om vijf uur wakker
Machine Translation
_Wa/ke/_me/_up/_at/_five/_am
_Wake/_me/_up/_at/_f/ive/_am
_Wa/ke/_me/_up/_at/_fiv/e/_am
Subword Sampling
Wake me up at five am
Input Utterance

Figure 2: Illustration of our fine-tuning framework. 'MT' denotes machine translation augmentation and 'SS' denotes subword sampling augmentation.

### 3.2.1 Subword Sampling

Subword sampling is to generate multiple subword sequences from the original text as data augmentation. We apply the on-the-fly subword sampling algorithm from the unigram language model (Kudo, 2018) in SentencePiece (Kudo and Richardson, 2018). The output distributions of slot labels are generated on the first subword of each word in the input utterance. Therefore, the subword sampling augmentation can be used to align the output distribution of both intent detection and slot filling tasks.

### 3.2.2 Machine Translation

Machine translation is a common and effective data augmentation strategy in the cross-lingual scenario (Conneau and Lample, 2019; Singh et al., 2019). Due to the difficulty of accessing ground-truth labels in translation examples, machine translation can not be an available data augmentation strategy in the slot filling task. To improve the quality of our translations, we employ a variety of approaches (See Section 4.2). Unlike subword sampling, the output distributions of slot labels between the translation pairs can not be aligned. Thus, we only use machine translation to align the output distributions of the intent detection task.

## 3.3 Consistency Regularization based on Hybrid Data Augmentations

We illustrate our fine-tuning framework in Figure 2. We propose to use consistency regularization based on a hybrid data augmentation strategy, which includes data augmentation of machine translation and subword sampling. During the training process, we perform task fine-tuning and consistency regularization for an input example simultaneously. Then the final training loss is defined as follows:

$$\mathcal{L} = \mathcal{L}_I + \lambda_1 \mathcal{L}_S + \lambda_2 \mathcal{R}_I + \lambda_3 \mathcal{R}_S$$

where $\lambda_1$ is the slot loss coefficient, $\lambda_2$ and $\lambda_3$ are the corresponding weights of the consistency regularization for two tasks. We sample different data augmentation for the input example with the pre-defined distribution.

# 4 Experiments

## 4.1 Experimental Setup

We consider two types of pre-trained cross-lingual language models, which are encoder-only models and Text-to-Text models.

We use XLM-Align Base (Chi et al., 2021b) for the encoder-only model setting. We use a two-layer feed-forward network with a 3,072 hidden size. We use the first representation of sentences "<s>" for the intent detection task and the first subword of each word for the slot filling task.

We use mT5 Base (Xue et al., 2021) for the Text-to-Text model setting. We follow FitzGerald et al. (2022) to concatenate "Annotate: " and the unlabeled input utterance as the input of the encoder, and generate the text concatenation of the intent label and the slot labels as the decoder output. The labels are separated with white spaces and then tokenized into subwords.

We select the model that performs the best on the development dataset to run prediction on the test and evaluation dataset. We mainly select the batch size in $[32, 64, 128, 256]$, dropout rate in

| Text Type | Text Content | Slot Translation | Text Translation | Aligned or Not |
|---|---|---|---|---|
| Plain Text | Wake me up at five am Friday this week | five am: 凌晨五点 | 本周周五凌晨五点叫我起床 | Yes |
| Text with Slots in Brackets | Wake me up at [five am] [Friday this week] | Friday this week: 本周周五 | 在[凌晨五点][本周星期五]叫醒我 | No |
| Plain Text | set an alarm for two hours from now | two hours from now: | 从现在开始设置两个小时的闹钟 | No |
| Text with Slots in Brackets | set an alarm for [two hours from now] | 从现在起两小时后 | 设置[从现在起两小时后]的闹钟 | Yes |

Table 1: Examples of aligning slots into machine translations.

| Model | Test Set | | | Evaluation Set | | |
|---|---|---|---|---|---|---|
| | Intent Acc | Slot F1 | EMA | Intent Acc | Slot F1 | EMA |
| XLM-R Base | 85.10 | 73.60 | 63.69 | - | - | - |
| XLM-Align Base | **86.16** | 76.36 | 66.42 | - | - | - |
| mT5 Base Text-to-Text | 85.33 | **76.77** | **66.64** | - | - | - |
| XLM-Align Base + Ours | 87.12 | 77.99 | 68.76 | 85.00 | 68.45 | 48.64 |
| mT5 Base Text-to-Text + Ours | **87.60** | **78.22** | **69.60** | **85.10** | **69.08** | **49.65** |

Table 2: Test and evaluation results on the MASSIVE dataset under the full-dataset setting. Results of XLM-R Base and mT5 Base Text-to-Text are taken from FitzGerald et al. (2022).

$[0.05, 0.1, 0.15]$, and the hyper-parameters used in our proposed method, including slot loss coefficient $\lambda_1$ in $[1, 2, 4]$, weights of consistency regularization $\lambda_2$ and $\lambda_3$ in $[2, 3, 5, 10]$. We select the learning rate in $[5e^{-5}, 8e^{-5}, 1e^{-4}]$ for Text-to-Text models. As for encoder-only models, we select the learning rate in $[4e^{-6}, 6e^{-6}, 8e^{-6}]$.

### 4.2 Data Processing

For the full-dataset setting, we use examples with the same id in different languages as machine translation augmentation in our fine-tuning framework. For the zero-shot setting, we translated the entire English training set into 50 languages using commercial translation APIs, such as DeepL translator and Google translator. These translations refer to plain text translations and can be used for intent detection training and consistency regularization.

We used two methods to obtain a translated example that aligned at the slot level. One is based on the plain text translation. Each slot value in an English training example is translated into a target language. If the translation results of each slot can be found in the plain text translation, a slot-aligned translation is obtained. The other is based on the annotated English training examples. We translated the annotated English training example with brackets for slot values (without slot type in brackets). Using brackets explicitly allows the translator to align slots to consecutive spans. And we also translated each slot value into the target language. If the translation result of each slot can be found in the annotated utterance translation, we obtain a slot alignment example after removing the brackets.

In practice, slot-aligned examples based on plain text translations are preferred as the final result of the slot alignment. If no such example is available, we use the slot-aligned results from annotated translations. Examples of slot alignment are shown in Table 1. For those plain text translations where we can not align the slot labels, we only use them for the training of the intent detection task.

### 4.3 Evaluation Metrics

The evaluation in competition is mainly conducted using three metrics:

- Exact Match Accuracy (EMA): The percentage of utterance-level predictions where the intent and all slots are exactly correct.
- Intent Accuracy (Intent Acc): The percentage of predictions in which the intent is correct.
- Slot Micro F1 (Slot F1): The micro-averaged F1 score is calculated over all slots.

### 4.4 Results

Table 2 shows our results on the MASSIVE dataset under the full-set setting. We tried different cross-lingual pre-trained language models under the baseline setting. Among them, XLM-Align Base performs the best on the intent detection task, while the mT5 Base Text-to-Text model performs the best on the slot filling task and exact match accuracy. When applying our consistency regularization method, the mT5 Base Text-to-Text model outperforms the XLM-Align Base model by 0.84 points and 0.99 points on exact match accuracy on the test dataset and the evaluation set, respectively. Meanwhile, compared to the baseline model, using consistency regularization achieves an absolute

| Model | Test Set | | | Evaluation Set | | |
|---|---|---|---|---|---|---|
| | Intent Acc | Slot F1 | EMA | Intent Acc | Slot F1 | EMA |
| XLM-R Base | 70.62 | 50.27 | 38.70 | - | - | - |
| XLM-Align Base | **68.49** | **54.69** | **40.91** | - | - | - |
| mT5 Base Text-to-Text | 62.92 | 44.77 | 34.72 | - | - | - |
| XLM-Align Base + Ours | 85.12 | 71.27 | 62.18 | 83.18 | 62.84 | 43.05 |
| XLM-Align Base + Ours + KD | **85.76** | **73.55** | **64.44** | **83.89** | **64.60** | **44.84** |
| mT5 Base Text-to-Text + Ours | 84.58 | 69.24 | 60.59 | 82.56 | 60.00 | 40.93 |

Table 3: Test and evaluation results on the MASSIVE dataset under the zero-shot setting. Results of XLM-R Base and mT5 Base Text-to-Text are taken from FitzGerald et al. (2022).

| Model | Intent Acc | Slot F1 | EMA |
|---|---|---|---|
| XLM-Align Base + Ours | 87.12 | **77.99** | **68.76** |
| - Subword Sampling | **87.50** | 76.08 | 67.40 |
| - Consistency Regularization | 86.16 | 76.32 | 66.57 |

Table 4: Ablation studies on the MASSIVE test dataset under the full-dataset setting.

| Model | Intent Acc | Slot F1 | EMA |
|---|---|---|---|
| XLM-Align Base + Ours | 85.12 | **71.27** | **62.18** |
| - Subword Sampling | **85.14** | 69.52 | 60.94 |
| - Machine Translation | 72.27 | 58.37 | 45.50 |
| - Consistency Regularization | 83.90 | 69.37 | 59.95 |

Table 5: Ablation studies on the MASSIVE test dataset under the zero-shot setting.

2.96-point improvement on exact match accuracy with the mT5 Base Text-to-Text model.

Table 3 shows our results on the MASSIVE dataset under the zero-shot setting. For the baseline models, XLM-Align Base performs the best on all three metrics. Difference from the full-dataset setting, mT5 Base Text-to-Text models perform poorly under the zero-shot setting. We attribute it to the fact that Text-to-Text models strongly rely on the training data quality since most of the training data under the zero-shot setting are obtained with machine translation systems. When applying our consistency regularization method, the XLM-Align Base model outperforms the baseline model by 21.27 points. Distilled from the InfoXLM Large (Chi et al., 2021a) model will further improve the performance by an absolute 2.26-point.

### 4.5 Ablation Studies

We conduct ablation studies on the test dataset of MASSIVE under the two settings. Table 4 shows the results under the full-dataset setting. Ablating subword sampling will degrade the performance by 1.36 points on the exact match accuracy, where the performance drop comes mainly from the slot filling task, indicating the subword sampling augmentation mainly works on slot filling. Ablating consistency regularization will degrade the performance by 2.19 points on the exact match accuracy. The performances on both intent detection and slot filling tasks are decreased.

The zero-shot setting results are presented in Table 5. It can be observed that when machine translation augmentation is removed, the exact match accuracy drops by 16.68 points, while the performance on intent detection and slot filling are also significantly worse. We also removed the subword sampling augmentation, and the performance is found to have the same trend as in the full-dataset setting. An absolute 1.24-point drop on the exact match accuracy and an absolute 1.75-point drop on slot micro F1 demonstrate that subword sampling is more beneficial for the slot filling task. By removing the consistency regularization, the performance of exact match accuracy will degrade by 2.23 points. The performance shows a significant performance drop on both intent detection and slot filling tasks.

## 5 Conclusion

We propose to use consistency regularization based on a hybrid data augmentation strategy to improve the performance of multilingual SLU. The proposed method is flexible and can be easily plugged into the fine-tuning process of both the encoder-only model and the Text-to-Text model. The experimental results demonstrate the importance of consistency regularization and the hybrid data augmentation strategy, respectively.

## Acknowledgments

This work was supported by the National Key RD Program of China via grant 2020AAA0106501 and the National Natural Science Foundation of China (NSFC) via grant 62236004 and 61976072.

## References

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021a. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021b. Improving pretrained cross-lingual language models via self-labeled word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3418–3430, Online. Association for Computational Linguistics.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2022. XLM-E: cross-lingual language model pre-training via ELECTRA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6170–6182. 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. 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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067. Curran Associates, Inc.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2177–2190. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 66–75. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial training for large neural language models. *CoRR*, abs/2004.08994.

Libo Qin, Qiguang Chen, Tianbao Xie, Qixin Li, Jian-Guang Lou, Wanxiang Che, and Min-Yen Kan. 2022. GL-CLeF: A global–local contrastive learning framework for cross-lingual spoken language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2677–2686, Dublin, Ireland. Association for Computational Linguistics.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. CoSDA-ML: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3853–3860. ijcai.org.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. XLDA: cross-lingual data augmentation for natural language inference and question answering. *CoRR*, abs/1905.11471.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008. Curran Associates, Inc.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Bo Zheng, Li Dong, Shaohan Huang, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021a. Allocating large vocabulary capacity for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3203–3215. Association for Computational Linguistics.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021b. Consistency regularization for cross-lingual fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3403–3417, Online. Association for Computational Linguistics.