C5L7: A Zero-Shot Algorithm for Intent and Slot Detection in Multilingual Task Oriented Languages

Jiun-Hao Jhan[1], Qingxiaoyang Zhu[2], Nehal Bengre[3], and Tapas Kanungo[3]

[1]Carnegie Mellon University Silicon Valley, CA, USA
[2]University of California Davis, CA, USA
[3]Samsung Research America, CA, USA
[1]*jiunhaoj@andrew.cmu.edu*
[2]*qinzhu@ucdavis.edu*
[3]*{n.bengre, tapas.k}@samsung.com*

Abstract

Voice assistants are becoming central to our lives. The convenience of using voice assistants to do simple tasks has created an industry for voice-enabled devices like TVs, thermostats, air conditioners, etc. It has also improved the quality of life of elders by making the world more accessible. Voice assistants engage in task-oriented dialogues using machine-learned language understanding models. However, training deep-learned models take a lot of training data, which is time-consuming and expensive. Furthermore, it is even more problematic if we want the voice assistant to understand hundreds of languages. In this paper, we present a zero-shot deep learning algorithm that uses only the English part of the Massive dataset and achieves a high level of accuracy across 51 languages. The algorithm uses delexicalized translation to generate a multilingual parallel corpus with intent and slot labels for data augmentation. The training data is further weighted to improve the accuracy of the worst-performing languages. We report on our experiments with code-switching, word order, multilingual ensemble methods and other techniques and their impact on overall accuracy.

1 Introduction

Task-oriented languages have become standard in voice-enabled devices and voice assistants. While there has been extensive research on task-oriented dialogue systems in limited domains, most of these systems are built in a limited set of languages due to a lack of labeled multilingual corpus. Amazon's MASSIVE dataset is a new resource for task-oriented language understanding that has 996K utterances annotated with intent and slot labels, along with their translations into 51 languages. The MASSIVE dataset is a unique resource for conducting multilingual language understanding research, and in particular building zero-shot learning algorithms where using only one language data, the trained system can perform language understanding tasks in the rest of the unseen languages. The importance of such training algorithms cannot be understated – labeled data is expensive and time-consuming to generate and hence any approach that reduces the cost and time to train such a multilingual system is desirable.

There are numerous hurdles in creating a zero-shot multilingual language understanding system. While machine translation systems can be used for translating utterances and creating a parallel corpus for training, aligning slot labels across languages can be challenging. In addition, if we expect the multilingual model representation to leverage information across languages, the input text representation needs to have the same tokenization process across languages. Furthermore, low-density languages are hard to get open-source resources for.

In this paper, we first review the related work in Section 2. Next, we address the issues listed above by introducing a novel delexicalized annotated utterance translation algorithm that is described in Section 3. To align code representations across languages, we randomly switched the language for a small percentage of the words. Finally, we explored the possible impact of using all the utterance translations instead of just one utterance in a specific language and were surprised by the accuracy boost. These and other experiments and analyses are described in Section 4.

2 Related Work

Transformer-based large multilingual masked language models, such as mBERT (Devlin et al., 2018), XLM (Lam- ple and Conneau, 2019) XLM-R (Conneau et al., 2019; Goyal et al., 2021), and mT5(Xue et al., 2020), have prevailed in cross-lingual language understanding. These models are pre-trained on a large multilingual text corpus to create a language representation that allows cross-lingual transfer on down-streaming tasks, such as cross-lingual document classification (Schwenk and Li, 2018; Pappas and Popescu-Belis, 2017),

Proceedings of the Massively Multilingual Natural Language Understanding Workshop (MMNLU-22), pages 62 - 68

role labeling (Björkelund et al., 2009), question answering (Kwiatkowski et al., 2019; Chen et al., 2017; Lewis et al., 2019) and named entity recognition (Nothman et al., 2013; Al-Rfou et al., 2015). In the field of natural language understanding, Liu and Lane and Chen et al. trained for intent classification and slot-filling tasks jointly to learn the inherent correlation between the two tasks via multitask learning. Castellucci et al. further used a joint Bert-based model to detect intents and extract slots for the multi-lingual scenario including English and Italian languages. However, systematic work on massive languages datasets (51 locales including sufficient variance of language order types including subject-initial, verb-initial and no preferred word order) has not been paid enough attention until now due to the lack of labeled datasets.

With the availability of the MASSIVE dataset (FitzGerald et al., 2022) with annotation for slot-filling and intent classification, and virtual assistant evaluation metrics and scoring tools, we will able to push the state of the art of multilingual natural language processing for a task-oriented dialogue system (Razumovskaia et al., 2022; Tur et al., 2010). To tackle the difficulty/high cost of collecting low-resource language data previous work (Xu et al., 2020; Upadhyay et al., 2018; Schuster et al., 2018) explored the use of machine translation to get translated data (Wu et al., 2016) and utilize zero-shot learning (Palatucci et al., 2009) to transfer the understanding learned on one language to another language. However, the correspondence across all languages in terms of intent and slot alignment is insufficiently incorporated into the training and inference phases of the cross-lingual NLU model. In this paper, we explore how to represent connection among massive languages in the model. Besides, inspired by the common code-switch behavior and multilingual speakers and previous work on learning cross-lingual structure (Heredia and Altarriba, 2001; Wu et al., 2019; Auer, 2013), we further explore the use of code-switch and delexicalization as anchor points to bridge the transfer learning among languages.

3 Method

3.1 Data Augmentation

Generated Parallel Corpus To train a zero-shot learning model, using English data is not sufficient, as can be seen in the low baseline results in Table 1. To address the problem, we propose to utilize Google Translator to translate English data to the other 50 languages and create an annotated parallel corpus. While translating an utterance is simple, translating annotated utterances is difficult since the alignment of the slots like "time" and "date" is not always straightforward. Our solution is to delexicalize the slots in the given utterance (described in Section 3.2) and use the delexicalized utterance as input to Google Translator. Next, we create a lookup table to map the delexicalized slots and the slot values. We then translate the original slot values into the target language. Finally, we use the lookup table to substitute the translated slots values into the corresponding delexicalized tags in the translated utterance. This process results in a translated annotated utterance in the target language. Each annotated English utterance is thus translated into each of the 50 target languages while preserving the intent and slot annotations.

Augmentation for Low-Performing Languages Low-performing languages decrease the total performance dramatically. Augmenting data for low-

Method	Intent	Slot F1	Exact Match
Baseline	70.6 %	50.3 %	38.7 %
GPC	79.7 %	58.8 %	40.3 %
GPC+DE	81.1 %	58.84 %	40.3 %

Table 1: This table shows the comparison between using generated parallel corpus (GPC), delexicalization (DE) and not using delexicalization. Baseline results are for our implementation of Zero-shot Intent and Slot Prediction algorithm by FitzGerald et al. (2022). We see that augmenting the training data with the generated parallel corpus (GPC) gives us a significant boost to intent and slot accuracy. When we add delexicalized utterances in addition to GPC, (GPC + DE), we get a further boost to intent accuracy, but not much to slot accuracy.

Method	Full-Dataset		
	Intent	Slot F1	Exact Match
Baseline	85.10 %	73.60 %	63.70 %
BOS	85.72 %	75.01 %	65.12 %
BOS + LO	85.87 %	74.75 %	65.20 %

Table 2: Objective Functions Results. We evaluated three objective functions with the full dataset (instead of zero-shot learning). Baseline results are for our implementation of Intent and Slot Prediction algorithm by FitzGerald et al. (2022). BOS means the Bag of Slot and LO means the language word order prediction. These objective functions give slight improvement to slot accuracy.

Language Order	Order-Specific Models			All Language Model		
	Intent	Slot F1	Exact Match	Intent	Slot F1	Exact Match
ALL	-	-	-	85.66 %	75.12 %	65.35 %
SVO	86.23 %	74.54 %	64.84 %	86.18 %	74.65 %	65.00 %
SOV	75.69 %	63.67 %	50.53 %	85.11 %	74.50 %	64.60 %
VSO	66.55 %	64.69 %	43.20 %	84.00 %	72.43 %	62.14 %
Uncategorized	77.88 %	69.72 %	54.25 %	86.03 %	74.41 %	64.74 %
None	82.31 %	70.73 %	58.76 %	86.37 %	74.49 %	65.47 %

Table 3: Languages Word Order Results. Order-Specific Models mean the models are trained on a specific language word order class and evaluated in the same class. All Language Model is trained jointly with all languages and evaluated on a specific language word order. Using all languages improves accuracy.

performing languages is one possible approach to address this issue. We collect the lowest performing ten languages and reweight the data by 2x and 5x.

3.2 Code Switching

To align model representation across languages, researchers (Lee et al., 2019) have used the notion of "code-switching," where they randomly switch the language of a small percentage of the words in the training corpus. We used a similar approach in our model training process. We identified common stop words across languages and used their English translations for random code-switching. For non-space separated languages ("zh-CN", "zh-TW", "ja-JP"), we do code-switching with 8%, while the rest languages are with 16% of the words. Code-switching potentially creates anchor points (the common sequences in different languages) across multiple languages and assists transfer learning.

3.3 Delexicalized Training Data

Earlier, we used slot delexicalization to generate the parallel multilingual corpus for training data augmentation. In this section, we use delexicalization for a different purpose. We use slot delexicalization to learn slot *usage patterns*. We delexicalized the slots randomly. The various slot values are replaced by slot types. For example, the annotated utterance "Wake me up at [time : five am] [date : this Friday]" is delexicalized to "Wake me up at TIME_SLOT DATE_SLOT". We delexicalize utterances in each language to learn shared features in the multilingual dataset. We delexicalize the input utterance slots with a probability of $\epsilon = 0.1$ while training.

3.4 Objective Functions

We represented the problem as a multi-task recognition problem. The models were initialized with a pre-trained XLM-Roberta (XLM-R) (Conneau et al., 2020) language model and fine-tuned it on the MASSIVE dataset (mas). We then trained four different classification heads from scratch: intent and slots prediction, bag of slot labels, and language order prediction in parallel.

Intent and Slot Prediction Our model is aimed to do intent classification and slot-filling tasks in the zero-shot scenario. We used the training process described in mas. We use the English subset of the data and augment it with our (generated) annotated parallel corpus as described in Section **??**. For intent classification, the model predicts the intent by using the pooled output from the XLM-R encoder which is the sentence-level embedding vector. Then, the model predicts slot logits (as a sequence labeling task) using XLM-R encoder representations of each token in the utterance. Then the CrossEntropy loss function is used to compare the intent and slot logits with ground truth labels to get the intent and slot loss.

Bag of Slot Labels (BOS) Since each utterance has 51 translated versions, we leverage the constraint that all 51 utterances have the same intent and slot labels. We batched the English utterance and the corresponding utterances in other languages into one block. The meaning of the utterances in the unit is the same. The only difference is that they are written in different languages. We expect the predictions within a unit to be as similar as possible. Thus, in this block of parallel multilingual utterances, ideally, each of the utterances should predict the same slot labels. (Although the slot labels across languages may not be aligned at each token, the set of *B-SLOTNAME* and *I-SLOTNAME* slot tags (in the BIO format) in each utterance inside a batch is the same as others. We represent the bag of slot labels as a D_{slots} dimensional binary vector with each location indicating which slots la-

bels are present in an utterance, where D_{slots} is the number of slot labels.) We collect 51 predictions as the output of intent classification and slot filling. Then we apply the CrossEntropy loss between the 51 intent predictions with ground truth.

Since the number of words in an utterance across the 51 languages and their word order might be different, computing loss per token does not work since the tokens are not aligned across languages. Thus, we get the mean of 51 languages' slot predictions and calculate the frequency of each slot type among these 51 utterances. Computing the CrossEntropy loss between the mean slot label predictions and the frequency might align the slot label predictions across the 51 predictions.

Language Word Order Prediction (LO) Word order is important in language. There are complicated rules for ordering words in different languages: two same utterances in different languages might generate large differences in the word's position in the sentence. Some languages start a sentence with the subject (S) following the verb (V) and the object (O). Others might start with the verb and end with the object. Therefore, we create another head to predict the language word order given an input utterance, training on the MASSIVE dataset. There are 5 kinds of word order in the MASSIVE dataset, SVO, SOV, VSO, none type, and uncategorized. We compute the CrossEntropy loss function between the order prediction and the ground truth. This loss function acts as one of the multitask among our objective functions.

4 Experiments and Results

4.1 Impact of Generated Parallel Corpus

The original baseline zero-shot algorithm described in mas uses only the English subset of the MASSIVE dataset and fine-tunes the multilingual XLMR model. We first explore the impact of augmenting the English subset of MASSIVE dataset with our generated (annotated) parallel corpus. In Table 1 we can see that our data augmentation increases the intent accuracy by 9.1% absolute and improves the average slot F1 score by 8.5%.

4.2 Augmenting Delexicalized Utterances

Next, in addition to augmenting the data with the generated parallel corpus, we added the delexicalized utterances. Table 1 shows that after applying the delexicalization technique, the intent accuracy increased by an absolute 2%. However, delexicalization barely improves the slot F1 score. The delexicalized data represents utterance templates, which the model learns, and perhaps helps with the intent accuracy. It is unclear why the slot accuracy was not impacted, perhaps a higher probability of delexicalization will help.

4.3 Objective Functions Comparison

In this experiment, we evaluate three objective functions by training on the full dataset from the MASSIVE (not zero-shot) training setup and testing on the corresponding test set, as shown in Table 2. The baseline results of the Intent and Slot prediction objective function are our implementation of (FitzGerald et al., 2022).

After including the Bag of Slot (BOS) objective function, the Slot F1 score increased by 2%. The main reason is that our model is capable of leveraging the shared information among 51 languages. However, adding the language word order prediction (LO) did not improve the performance. We found that the accuracy of language word order prediction is close to 100% and the loss is close to 0. The implication is that the XLMR model has learned to classify the language word order very well. However, the constraint of predicting language word order barely influenced the overall result.

4.4 Language Word Order Prediction Results

In this experiment, instead of training all languages jointly, we trained five different models corresponding to the language word order. In FitzGerald et al. (2022) the authors classified the language word order into five classes, SVO, SOV, VSO, Uncategorized, and None. According to results presented in Table 3, training a single SVO class model gets a similar performance as training on all languages jointly, while other classes get worse results. The main reason is that languages in the SVO class, like English, Spanish, etc., dominated the dataset. XLMR pretrained model is capable of understanding languages in the SVO class well. As for other language word orders, there might be a low-resource data problem while training the pretrained model that gives rise to a huge accuracy difference with respect to SVO class languages. In addition, training with all languages gives us better performance than training with only one language. The reason might be that training jointly makes the model leverage common characteristics amongst different language word orders.

Method	Full-Dataset				
	Training Method	Test Dataset	Intent	Slot F1	Exact Match
Amazon XLM-Base	full-training	MMNLU test	85.10 %	73.60 %	63.70 %
Amazon XLM-Base	zero-shot	MMNLU test	70.6 %	50.30 %	38.70 %
XLM-Base +BOS+DE	zero-shot	MMNLU-22 test	81.55 %	59.26 %	40.49 %
XLM-Base+GPC +BOS+DE+Ensemble	zero-shot	MMNLU-22 test	88.13 %	59.42 %	42.08 %

Table 4: Ensemble Result on MMNLU-22 Test Split. We evaluated our final model (the model with BOS and DE in training, Ensemble in post-processing) with MMNLU-22 test split, which is the test split of MMNLU-22 competition zero-shot track. The model was trained with the GPC dataset. The result of Amazon XLM-Base's model using full data training and the result of Amazon XLM-Base using zero-shot training on en-US are referred from the original MMNLU paper FitzGerald et al. (2022). BOS means the Bag of Slot, DE means delexicalized, and GPC means generated parallel corpus. The ensemble strategy gives significant improvement to intent accuracy on MMNLU-22 test set, making it even higher than Amazon's full-training dataset baseline results on MMNLU test set.

4.5 Post Processing with Ensemble Method

To leverage the characteristic of the parallel dataset, we experimented with an ensemble technique. Since for each utterance we have 50 translations with the same intent, we make each language vote for an intent and select the intent with the most votes as the final intent for all 51 languages. As a result (shown in Table 4), our model, including three objective functions and the voting technique, achieves 88.13%, 59.42%, and 42.08% for intent accuracy, slot F1 score, and exact match accuracy, respectively in MMNLU-22 test split[1]. In fact, intent accuracy achieves a significant boost with 6.61% in comparison to the result without ensemble strategy. We also see that the resulting intent accuracy is higher even than Amazon's baseline full-training data set. The slot F1 score, though significantly higher than Amazon's zero-shot baseline, is still much lower than the full-training data set results. This is probably due to using translations of English slot values to target languages. In our experiment, we used the translations from English to the target languages. However, to apply the voting technique in practice, we need to translate the utterance in the input language to all other target languages to elicit our model's multi-perspective on other languages and get a robust prediction through the ensemble. This work is currently in progress.

[1]The website of the competition with leaderboard: https://eval.ai/web/challenges/challenge-page/1697/leaderboard/4061

5 Discussion and Future Work

How good is the delexicalized slot translation? One approach to quantify this would be to generate an annotated translation from English to language i using Google translator and then translate it back to English and then compute a BLEU score.

Our zero-shot ensemble method using generated parallel corpus gives us better intent accuracy than the baseline full-set result in (FitzGerald et al., 2022). However, the slot accuracy is still much lower. One of the reasons could be that the slot values don't translate well to other languages. For example, a Christian name is not something that will be common in Chinese data. Using language-specific values probably will yield better results.

The ensemble method in a real-world setting requires us to translate utterance t in language i, to all other 50 languages. This requires to generate n^2 translation, which is expensive on Google Translate. For our experiment, we instead used the translations from English. One issue with this approach could be that English-to-target language translation might be of better quality than the translation of input language to a target language. Doing the full experiment will be conclusive. Another drawback of the ensemble approach is the need for n real-time translations and n parallel real-time runs. However, one way to reduce this complexity is to find a small subset of languages that we can use for voting purposes.

6 Conclusion

We presented a zero-shot, multilingual, joint intent-detection and slot-filling algorithm based on XLM-

R Transformer and Amazon's MASSIVE dataset. We showed that our delexicalized translation approach to generating a parallel corpus for data augmentation is a viable approach for training zero-shot algorithms. We showed that training using data from all language order types gives superior accuracy than using only a single language order type data in most cases – n MASSIVE data, the SVO category performed equally well when using just the SVO subset. Furthermore, our experiments showed that using an ensemble approach with translations of the input utterance can lead to a significant gain in accuracy.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments on an earlier draft of this paper and Steve Walsh and Yuri Lozhnevsky for their assistance.

References

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594. SIAM.

Peter Auer. 2013. *Code-switching in conversation: Language, interaction and identity*. Routledge.

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 43–48.

Giuseppe Castellucci, Valentina Bellomaria, Andrea Favalli, and Raniero Romagnoli. 2019. Multi-lingual intent detection and slot filling in a joint bert-based model. *arXiv preprint arXiv:1907.02884*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

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*.

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. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, et al. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *arXiv preprint arXiv:2204.08582*.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. *arXiv preprint arXiv:2105.00572*.

Roberto R Heredia and Jeanette Altarriba. 2001. Bilingual language mixing: Why do bilinguals code-switch? *Current Directions in Psychological Science*, 10(5):164–168.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Grandee Lee, Xianghu Yue, and Haizhou Li. 2019. Linguistically motivated parallel data augmentation for code-switch language modeling. In *Interspeech*, pages 3730–3734.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv:1609.01454*.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. 2009. Zero-shot learning with semantic output codes. *Advances in neural information processing systems*, 22.

Nikolaos Pappas and Andrei Popescu-Belis. 2017. Multilingual hierarchical attention networks for document classification. *arXiv preprint arXiv:1707.00896*.

Evgeniia Razumovskaia, Goran Glavas, Olga Majewska, Edoardo M Ponti, Anna Korhonen, and Ivan Vulic. 2022. Crossing the conversational chasm: A primer on natural language processing for multilingual task-oriented dialogue systems. *Journal of Artificial Intelligence Research*, 74:1351–1402.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2018. Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327*.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. *arXiv preprint arXiv:1805.09821*.

Gokhan Tur, Dilek Hakkani-Tür, and Larry Heck. 2010. What is left to be understood in atis? In *2010 IEEE Spoken Language Technology Workshop*, pages 19–24. IEEE.

Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038. IEEE.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Emerging cross-lingual structure in pretrained language models. *arXiv preprint arXiv:1911.01464*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual nlu. *arXiv preprint arXiv:2004.14353*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.