Zero-Shot Cross-Lingual Sequence Tagging as Seq2Seq Generation for Joint Intent Classification and Slot Filling

Fei Wang[*], **Kuan-Hao Huang**[*], **Anoop Kumar, Aram Galstyan, Greg Ver Steeg, Kai-Wei Chang**
Amazon Alexa AI

Abstract

The joint intent classification and slot filling task seeks to detect the intent of an utterance and extract its semantic concepts. In the zero-shot cross-lingual setting, a model is trained on a source language and then transferred to other target languages through multi-lingual representations without additional training data. While prior studies show that pre-trained multilingual sequence-to-sequence (Seq2Seq) models can facilitate zero-shot transfer, there is little understanding on how to design the output template for the joint prediction tasks. In this paper, we examine three aspects of the output template – (1) label mapping, (2) task dependency, and (3) word order. Experiments on the MASSIVE dataset consisting of 51 languages show that our output template significantly improves the performance of pre-trained cross-lingual language models.

1 Introduction

The joint intent classification and slot filling task is crucial for goal-oriented dialogue systems, seeking to detect the intent of an utterance and extract semantic concepts. This task has been widely studied in the literature (Hakkani-Tür et al., 2016; Zhang and Wang, 2016; Goo et al., 2018). However, due to the difficulty of collecting and annotating large data sets, most studies focus on only a few high-resource languages (e.g., English). To broaden the language coverage of models, zero-shot cross-lingual transfer technique has been proposed (Xu et al., 2020; Li et al., 2021; FitzGerald et al., 2022). Under the zero-shot cross-lingual setting, models are trained on a source language (e.g., English) with sufficient annotated training data and transfer to other target languages.

In particular, recently, FitzGerald et al. (2022) show that pre-trained generative cross-lingual language models (XLMs) (Liu et al., 2020; Xue et al., 2021) can be applied to the joint intent classification and slot filling task. They formulate the joint task as sequence-to-sequence (Seq2Seq) generation, where the model generates the slot label for each word and the intent label for the utterance in a sequential manner based on an output template. However, the design of the output template is usually ad hoc and there is lack of understanding on how different template designs affect the performance of the zero-shot transfer. For example, in Fig. 1, the intent label "`alarm set`" can be represented by "`set alarm`". We found that the change of the surface form of the label significantly affects model's performance.

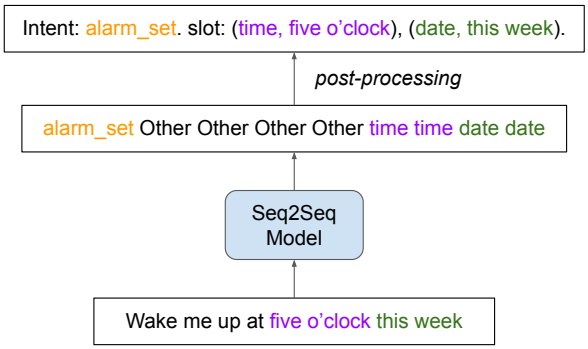

Figure 1: Illustration of Seq2Seq generation for the joint intent classification and slot filling task. Given an input on the bottom, the Seq2Seq model generates the output sequence based on a template – the template forces the model first output the intent label and then slot label of each word in the input sentence. Based on the template, a post-processing step translates the output sequence into structured labels for the task.

In this paper, we examine three aspects in the design of output template, i.e. label mapping, task dependency, and word order. We found that all these aspects have significant influence on model performance. First, based on our observation on label mapping, we propose a concise hierarchical label mapping that leads to a better performance than the default label mapping used in annotation. Second, we observed that generating the intent label before

[*]This work is done during Fei Wang and Kuan-Hao's internship at Amazon.

Proceedings of the Massively Multilingual Natural Language Understanding Workshop (MMNLU-22), pages 53 - 61

the slots labels leads to better performance. This is because intent classification is a relatively simpler task compared to slot filling and thus the correct task order prevents error propagation. Finally, we found that word shuffle improves the diversity of data and therefore leads to better intent accuracy.

Experiments on the MASSIVE dataset (FitzGerald et al., 2022) consisting of 51 languages demonstrate that our proposed template design can significantly improve the performance of pre-trained generative XLMs on this joint task in the zero-shot transfer setting. We also provide detailed ablation studies and discussion. We intend to release the source code for reproducing our experiments upon paper acceptance.

2 Method

We first provide an overview of the Seq2Seq generation method with pretrained generative XLMs for the joint intent classificaiton and slot filling task. Then, we discuss the design of output template for Seq2Seq generation.

2.1 Seq2Seq for the Joint Prediction Task

Following FitzGerald et al. (2022), we formulate the joint task as Seq2Seq generation, and adopt pretrained XLMs for this task. In this way, we can take advantage of the rich cross-lingual knowledge possessed in the pre-trained XLMs. As shown in Fig. 1, the model generates the intent label for the utterance and the slot label for each word in a sequential manner based on an output template. We insert `Annotate:` at the beginning of the input sequence to indicate the task type. We also insert word separators to indicate the tokens belonging to each word, as we want to generate word-level slot labels. We then use the following objective function to fine-tune the text generation model:

$$\mathcal{L} = -\frac{1}{|\mathbf{y}|}\sum_{k=1}^{|\mathbf{y}|}\log p(y_k|\mathbf{y}_{<k}, \mathbf{x}), \qquad (1)$$

where $\mathbf{x}$ is the input utterance sequence and $\mathbf{y}$ is the output label sequence. We provide the ground-truth $\mathbf{y}$ during training. In the following, we explore three aspects of the design of the output template.

2.2 Label Mapping

The first aspect we examine is the label mapping, i.e. the surface form of the labels. When annotating data, the annotators are given a set of output labels. The choice of vocabulary for these labels is often arbitrary as long as the human annotators understand the meaning. However, for a Seq2Seq model, different surface forms of the output labels may lead to different performance even though they are synonyms. For example, the intent label `iot_wemo_on` could be difficult for a fine-tuned XLMs to understand and transfer to other languages. Moreover, labels hold hierarchical relations. For example, some intent labels may belong to the same scenario and some slot labels belong to the same intent. By rephrasing the output labels and leveraging their relations, the model performance can greatly improve.

We propose a concise and hierarchical label mapping based on these observations. For labels belonging to the same scenario or intent, we add the same prefix to them. Some slot labels may belong to multiple scenarios or intents, so we do not add any prefix to them. For example, both `email_folder` and `email_address` belongs to the same scenario so we give them the same prefix `email`, while `time` belongs to multiple scenarios, so we do not give it a prefix. We also remove or replace the redundant and rare words in the labels (e.g. `wemo` in `iot_wemo_on`).

2.3 Task Dependency

The second aspect is the dependency between intent classification and slot filling. In Seq2Seq decoding, the label y_k is conditioned on previously generated tokens $\mathbf{y}_{<k}$. When solving the joint task, the label of one task serves as the condition to generate the label of the other task. Due to this, the later task may benefit from the labels of the former task, but may also suffer from inaccurate predictions of the former task (i.e. error propagation).In particular, we consider two different orders: (1) intent labels before slot labels, and (2) slot labels before intent labels.

2.4 Word Order

Prior works show that reducing word order information in sequence labeling can improve cross-lingual transferability (Ahmad et al., 2019; Liu et al., 2021). This is mainly due to different languages have different word orders (e.g., some languages present adjectives before nouns and some have reverse order), which cause a misalignment in language transfer. In Seq2Seq decoding, changing word orders results in different label order. To make the model more robust on different word orders, we augment the training data by shuffling the utterances and

their corresponding labels. However, different from prior works, we shuffle the utterances at the segment level, where words belonging to the same slot is considered as one segment and adjacent words that do not belong to any slot are considered another segment.

3 Experiment

In this section, we evaluate our approach on a massive number of target languages.

3.1 Setup

Dataset. We adopt the MASSIVE (FitzGerald et al., 2022) dataset, which consists of 51 languages, 18 domains, 60 intents, 55 slots and 19,521 utterance per language. We use English data for training and development, data in all the other languages are used for testing.[1] We report the intent accuracy, micro-average slot F1 and exact match accuracy.[2]

Baseline. We compare our method with both classification based and generation based methods. The classification method based on XLM-R (Conneau et al., 2020) formulates the joint task as sequence classification and sequence tagging. Two classification heads are added on top of the pre-trained language model. The generation method based on mT5 (Xue et al., 2021) generates the tag of each word and the intent label in a sequence-to-sequence manner. Following FitzGerald et al. (2022), we use the base version of pre-trained models.

Implementation Details. We evaluate our method based on mT5 with the original model-related hyper-parameters. We follow the hyper-parameters of FitzGerald et al. (2022) for training, except for batch size, learning rate and epochs, which we set to 96, 5e-5 and 200, respectively. We investigate three design choices listed in Sec. 2. We found that better label mapping and task dependency significantly improves the model performance. However, while input shuffle improves intent accuracy, the slot F1 and exact match performance drop. In the following, we will first compare our best model (w/ label mapping and w/ task dependency) with the current state-of-the-art approach, then we will provide detailed ablation study.

[1]This setting is more strict than FitzGerald et al. (2022)'s, where they use data in target languages for development.

[2]Exact match means both the intent and slots are correct.

3.2 Results

Tab. 1 shows the overall model performance on MASSIVE. In comparison with the vanilla mT5, our proposed techniques improve average intent accuracy by 1.5%, average slot F1 by 10.7% and average exact match accuracy by 5.2%. It also changes the highest performing languages in terms of slot F1 and exact match accuracy. These results indicate that task order, label mapping, and other key components in text generation have significant influence on model performance when performing sequence tagging in a sequence-to-sequence generation manner. The key differences between XLM-R based and mT5 based methods are that the latter ones use pre-trained token embeddings as labels and generate each label conditioned on previously generated labels. The vanilla mT5 performs much worse than the prior SOTA method, XLM-R, on all metrics. However, our method based on mT5 achieves better performance than XLM-R in terms of average slot F1 (+5.0%) and exact match accuracy (+2.2%). The failure of vanilla mT5 further shows the importance of well-designed inputs and outputs. The lowest performing language is consistent to be Japanese in all methods. Our method improves slot F1 and exact match accuracy of the lowest performing language.

Performance on Language Characteristics. We further analyze the model performance on different language characteristics. As shown in Fig. 2, our method performs better than vanilla mT5 on 49 out of 50 languages, indicating it can improve the cross-lingual transferability on massive target languages. Norwegian is the only language on which our method performs slightly worse. We provide detailed model performance on 9 language characteristics in Appx. §A, where the languages are split into 3 to 28 groups by each characteristic. Our method improves the performance of all language groups, except for Lolo-Burmese subdivision and Burmese script which contain only Norwegian. We observe that it is difficult to improve model performance on the Japonic and Sino-Tibetan language families when using English as source language. Similarly, a prior work (Malkin et al., 2022) also shows that English may not be an optimal pretraining language in cross-lingual transfer. We leave finding the best general source language for finetuning zero-shot models for future work.

Ablation Study. We also investigate the effective-

Model	Intent Acc (%)			Slot F1 (%)			Exact Match Acc (%)		
	High	Low	Avg	High	Low	Avg	High	Low	Avg
mT5	79.9 ± 1.4 nl-NL	25.7 ± 1.6 ja-JP	62.9 ± 0.2	64.3 ± 0.7 de-DE	13.9 ± 0.3 ja-JP	44.8 ± 0.1	53.2 ± 1.8 sv-SE	9.4 ± 1.0 ja-JP	34.7 ± 0.2
XLM-R	85.2 ± 1.3 sv-SE	44.8 ± 1.8 ja-JP	**70.6** ± 0.2	68.4 ± 0.7 sv-SE	15.4 ± 0.3 ja-JP	50.3 ± 0.1	57.9 ± 1.8 sv-SE	9.8 ± 1.1 ja-JP	38.7 ± 0.2
mT5*	80.6 ± 0.7 nl-NL	32.1 ± 0.9 ja-JP	64.8 ± 0.1	63.9 ± 0.3 de-DE	14.7 ± 0.2 ja-JP	44.6 ± 0.1	54.1 ± 0.9 sv-SE	10.1 ± 0.6 ja-JP	35.7 ± 0.1
OURS	80.8 ± 0.7 nl-NL	24.6 ± 0.8 ja-JP	66.3 ± 0.1	71.6 ± 0.5 th-TH	19.6 ± 0.2 ja-JP	**55.3** ± 0.1	57.4 ± 0.9 nl-NL	10.2 ± 0.5 ja-JP	**40.9** ± 0.1

Table 1: Zero-shot cross-lingual results on MASSIVE. We report intent accuracy, micro-averaged slot F1 score, and exact match accuracy of highest language, lowest language and average of all target languages. Best average scores are in bold. Intervals for 95% confidence are given assuming normal distributions. *: results reproduced by us. Other baseline results are copied from the original paper.

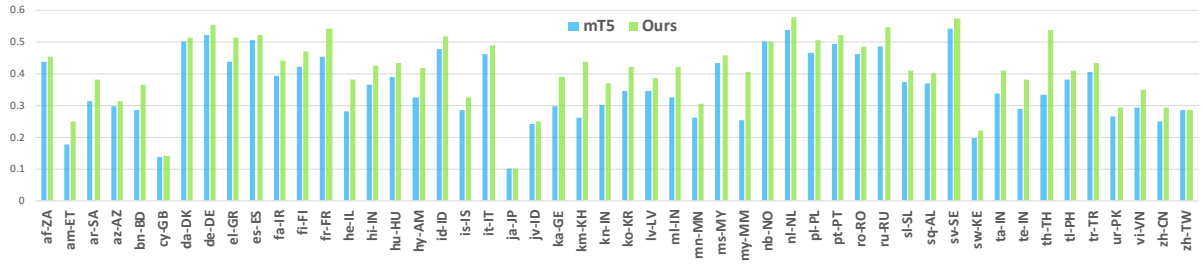

Figure 2: Exact match accuracy of all target languages. Our method performs better than vanilla mT5 on 49 out of 50 languages.

Method	Intent Acc	Slot F1	Exact Match
OURS	66.3	55.3	40.9
OURS w/ default label	67.0	52.5	39.7
OURS w/ slot first	64.6	57.5	37.1
OURS w/ input shuffle	67.5	52.6	40.1

Table 2: Ablation study. All intervals for 95% confidence are within ± 0.1%.

ness of each proposed technique as shown in Tab. 2. Input shuffle increases the diversity of inputs and help the model to avoid overfitting English syntax. Results show that it can improve the intent accuracy (+1.2%); however, it hinders predicting slots accurately. Concise and hierarchical label mapping improves the slot F1 significantly (+2.8%). Task order also plays an important role. Generating slot labels before the intent label for each utterance leads to worse intent accuracy (-1.7%) but better slot F1 (+2.2%). We observe the subtask performance is better when the model generates the subtask labels first.

4 Related Work

Zero-shot cross-lingual joint intent classification and slot filling is crucial for developing goal-oriented dialogue systems for massive languages with less manually annotation (Upadhyay et al., 2018; Li et al., 2021; FitzGerald et al., 2022). Prior works on this joint task can be summarized into two lines. The first line follows a strict zero-shot setting, where only the data in source languages are used for training (Xu et al., 2020; Li et al., 2021; FitzGerald et al., 2022). The second line uses additional data consisting of words or utterances in target languages for training, where the additional data can be annotated data in target languages or synthetic data by code-switching and automatic translation (Upadhyay et al., 2018; Schuster et al., 2019; Krishnan et al., 2021).

Our work follows the strict zero-shot setting. Prior works either formulate the joint task as sequence tagging and applies pretrained cross-lingual encoders (Pires et al., 2019; Conneau et al., 2020) to solve it (Xu et al., 2020; Li et al., 2021; FitzGerald et al., 2022), or formulate it as Seq2Seq generation and applies pretrained generative XLMs to solve it (FitzGerald et al., 2022). Our work analyzes important variables in the output format of the Seq2Seq method.

5 Conclusion

In this paper, we examine three variables of output format in Seq2Seq generation for zero-shot cross-lingual joint intent classification and slot filling. Experiments on the MASSIVE dataset consisting of 51 languages show that all the variables have significant influence on model performance. Specifically, the output format should use a concise and hierarchical label mapping, and consider the label dependency carefully.

Acknowledgement

The authors thank the anonymous reviewers for their valuable comments. They also thank I-Hung Hsu and Muhao Chen for their feedback.

Limitation

In this paper, we analyze three aspects of the design of output format in Seq2Seq generation for zero-shot transfer. There are other factors (e.g., decoding strategy) may also influence the model performance and its transferablity. Besides, this paper focuses on the *output* template of Seq2Seq generation in the cross-lingual transfer setting. We do not consider and compare with other techniques such as data augmentation methods for zero-shot cross-lingual transfer (e.g., code-switching (Qin et al., 2021) and robust training (Huang et al., 2021)).

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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 2440–2452.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.

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*.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757.

Dilek Hakkani-Tür, Gokhan Tur, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-Domain Joint Semantic Frame Parsing Using Bi-Directional RNN-LSTM. In *Proc. Interspeech 2016*, pages 715–719.

Kuan-Hao Huang, Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Improving zero-shot cross-lingual transfer learning via robust training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1684–1697.

Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. 2021. Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 211–223.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zihan Liu, Genta I Winata, Samuel Cahyawijaya, Andrea Madotto, Zhaojiang Lin, and Pascale Fung. 2021. On the importance of word order information in cross-lingual sequence labeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13461–13469.

Dan Malkin, Tomasz Limisiewicz, and Gabriel Stanovsky. 2022. A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4903–4915, Seattle, United States. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2021. Cosda-ml: multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3853–3860.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805.

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.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual nlu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063.

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*, pages 483–498.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2993–2999.

A Model performance on language characteristics

We compare the performance of vanilla mT5 and our method on different language characteristics, including script (Fig. 3), subdivision (Fig. 4), family (Fig. 5), order (Fig. 6), politeness (Fig. 7), imperative morphology (Fig. 8), imperative hortative (Fig. 9), optative (Fig. 10) and prohibitive (Fig. 11).

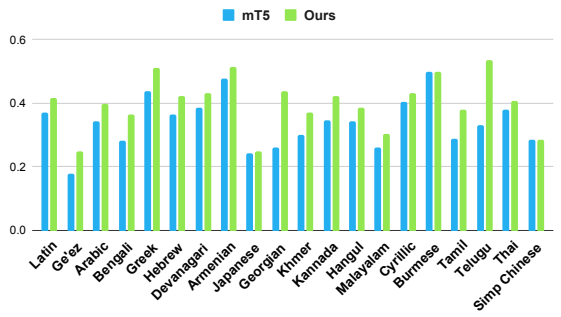

Figure 3: Exact match accuracy by language script.

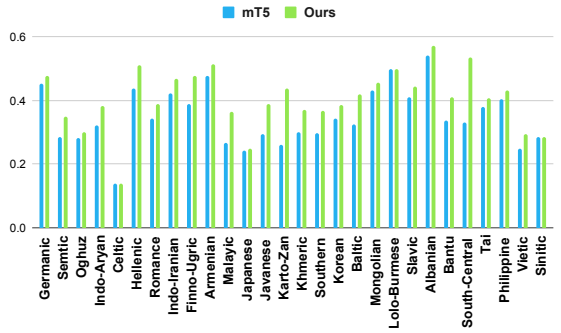

Figure 4: Exact match accuracy by language subdivision.

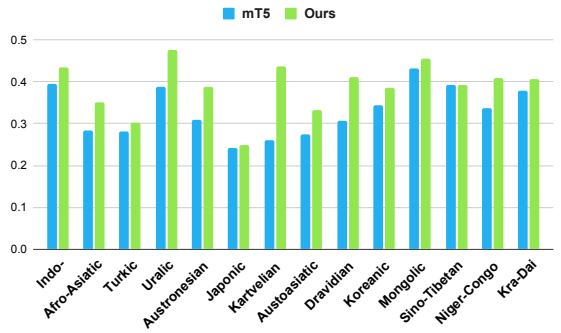

Figure 5: Exact match accuracy by language family.

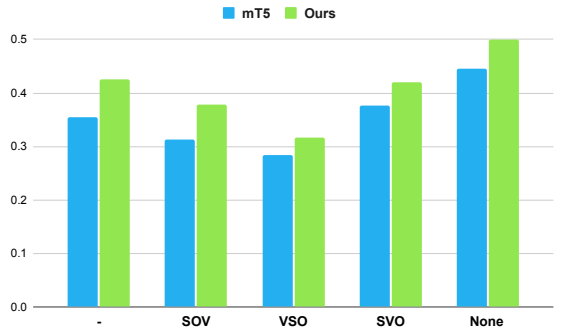

Figure 6: Exact match accuracy by language order.

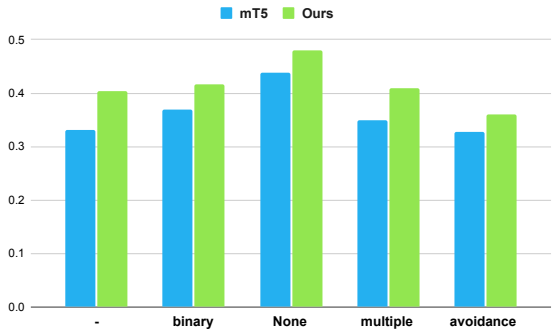

Figure 7: Exact match accuracy by language politeness.

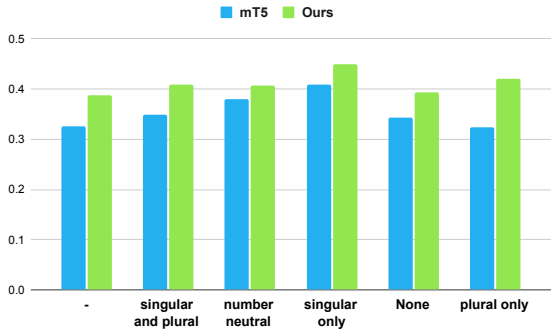

Figure 8: Exact match accuracy by language imperative morphology.

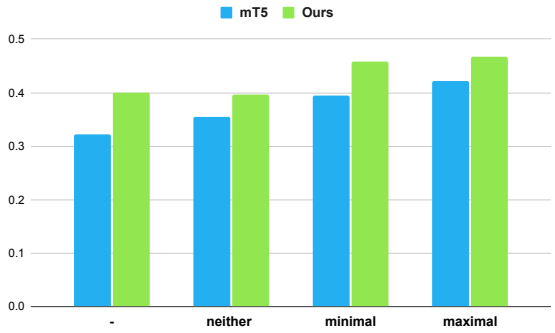

Figure 9: Exact match accuracy by language imperative hortative.

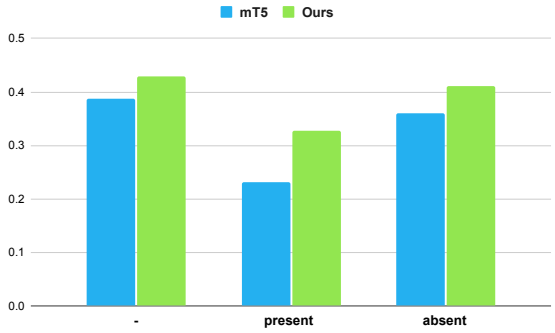

Figure 10: Exact match accuracy by language optative.

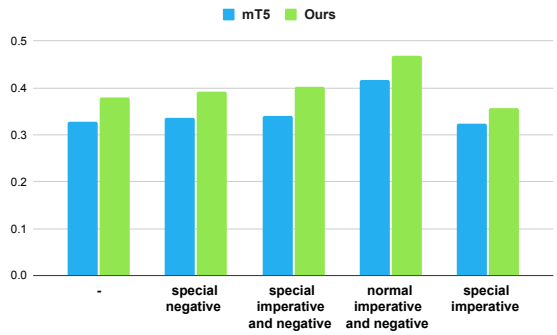

Figure 11: Exact match accuracy by language prohibitive.