Evaluating Byte and Wordpiece Level Models for Massively Multilingual Semantic Parsing

Massimo Nicosia and Francesco Piccinno
Google Research, Zürich
{massimon,piccinno}@google.com

Abstract

Token free approaches have been successfully applied to a series of word and span level tasks. In this work, we compare a byte-level (ByT5) and a wordpiece based (mT5) sequence to sequence model on the 51 languages of the MASSIVE multilingual semantic parsing dataset. We examine multiple experimental settings: (i) zero-shot, (ii) full gold data and (iii) zero-shot with synthetic data. By leveraging a state-of-the-art label projection method for machine translated examples, we are able to reduce the gap in exact match accuracy to only 5 points with respect to a model trained on gold data from all the languages. We additionally provide insights on the cross-lingual transfer of ByT5 and show how the model compares with respect to mT5 across all parameter sizes.

1 Introduction

Semantic parsers map natural languages utterances into logical forms (LFs). In the context of conversational agents (Artzi and Zettlemoyer, 2011), robotics (Dukes, 2014) or question answering systems (Berant et al., 2013), task-oriented semantic parsers map user queries (e.g. *"set an 8 am alarm"*) to machine readable LFs (e.g. `[IN:CREATE_ALARM [SL:TIME 8 am ]]`), in the form of structured interpretations that can be understood and executed by downstream components. Learning parsers requires training data in the form of <utterance, LF> pairs. Such data is costly to obtain especially at large scale (Berant et al., 2013), since expert annotators have to derive the correct LFs given an input utterance. This problem is exacerbated in a multilingual setting, where the availability of annotators, especially for non top-tier languages, is scarce and therefore even more expensive.

With the release of MASSIVE (FitzGerald et al., 2022), the research community has now access to a massively multilingual semantic parsing dataset that can be used to evaluate large language models fine-tuned on the task and to study cross-lingual transfer for numerous languages.

On the multilinguality front, token-free models with byte or character based vocabularies have gained strength given their competitiveness with respect to traditional subword-based pretrained language models. Models such as ByT5 (Xu et al., 2020), Canine (Clark et al., 2022) and the Charformer (Tay et al., 2022) have been applied to popular multilingual benchmarks obtaining state-of-the-art results.

In this paper, we perform the first in-depth evaluation of a token-free model in the context of multilingual semantic parsing. We compare the ByT5 and mT5 (Xue et al., 2021) models across different parameter sizes and data regime settings. In addition to that, we build a map of the cross-lingual transfer for all the languages in MASSIVE. Lastly, we show that with the use of machine translated synthetic data the accuracy of a state-of-the-art multilingual parser can be just 5 points lower than the same parser trained with all the available multilingual supervision. To incentivize research on synthetic data augmentation approaches, we release the MASSIVE English training utterances translated to 50 languages.[1]

2 The MASSIVE Dataset

MASSIVE (FitzGerald et al., 2022) is a semantic parsing dataset covering 51 languages, 18 domains, 60 intents and 55 slots. The dataset was created by professional translators starting from the English SLURP dataset (Bastianelli et al., 2020). A significant portion of the translations have been localized too, following the recent trend in multilingual benchmarks of replacing western-centric

[1] We release the translations in 50 languages of the MASSIVE English training examples obtained with an in-house translation system at `https://goo.gle/massive-translations`

Proceedings of the Massively Multilingual Natural Language Understanding Workshop (MMNLU-22), pages 25 - 34
December 7, 2022 ©2022 Association for Computational Linguistics

entities with entities that are more relevant for the target languages (Lin et al., 2021; Ding et al., 2022; Majewska et al., 2022).

2.1 Pre and Post Processing

The annotated instances in the MASSIVE dataset come in the following format:

```
intent: alarm_set
annot_utt: despiértame a las [time :
    ↪ nueve de la mañana] el [date :
    ↪ viernes]
```

To shorten the target output and save the model from generating and potentially hallucinating unnecessary words, we map the former to the following format taken from MTOP (Li et al., 2021):

```
[IN:ALARM_SET [SL:TIME nueve de la mañ
    ↪ ana ] [SL:DATE viernes ] ]
```

For evaluation, we use a simple inverse post-processing step based on string matching to convert the model outputs back to MASSIVE format.

2.2 Synthetic Data with Translate-and-Fill

A common approach to create multilingual synthetic data from available examples is to use machine translation (Moradshahi et al., 2020; Sherborne et al., 2020). Utterances are translated and LF annotations are projected using word aligners and noise reduction heuristics. We instead adopt the approach from Nicosia et al. (2021), Translate-and-Fill (TAF), a label projection method in which a filler model reconstructs the full LF starting from an utterance and its LF signature.

We train an mT5-xxl filler model on English instances and then directly generate the LFs of translated examples in a zero-shot fashion. Since the slot order between English and translated utterances may differ, we canonicalize the generated synthetic interpretations reordering the slots as they would occur in the translations. We have also noticed in the filler output that for some languages the slot boundaries may fall inside words. For languages with white space tokenization, we move slot boundaries to word boundaries if needed.

As an example, given an input utterance *"despiértame a las nueve el viernes"* and `[IN:ALARM_SET [SL:DATE el vier ] [SL:TIME nueve ] ]` as LF, the process looks as follows. First the arguments are reordered according to the order of appearance in the original sentence: `[IN:ALARM_SET [SL:TIME nueve ] [SL:DATE vier ] ]`. Then slot boundaries that fall within words are extended, correcting the prediction for the second argument from `[SL:DATE vier ]` to `[SL:DATE viernes ]`.

3 Experiments

We use MASSIVE as a test bed for two model families, ByT5 and mT5, evaluating them at all sizes in three different data settings. We report *Intent Accuracy* (IA) and *Exact Match* (EM) accuracy. We do not perform any hyper-parameter tuning: we train for 30K steps with a fixed learning rate of 0.0001 and a batch size of 128 for all models but xxl, for which batch size was reduced to 32. We run fine tuning on Cloud TPU v3 with an input/-target length of 1024/512 for ByT5 and 512/512 for mT5. To minimize compute, all the reported results are from single runs. We experiment with three different settings, summarized below:

1. **Zero-shot setting.** Training is performed on English data only, and the model selection is done on the English development set. Results are reported in Table 1.

2. **Gold-data setting.** Training is performed on all the MASSIVE data, that includes 51 languages. Model selection is performed averaging the accuracy on the multilingual development sets. Results are reported Table 2.

3. **Synthetic data setting (TAF).** Training is performed on English and multilingual data that is synthetically generated via TAF. Results are reported in Table 3. Our entry based on this approach ranked 1st in the Zero-Shot Task of the MMNLU-22 Multilingual Semantic Parsing competition organized by Amazon and co-located with EMNLP 2022.[2]

We can see a pattern that is common to all the experiments: at smaller sizes, ByT5 has much better EM accuracy then the corresponding mT5 models. As stated in Xu et al. (2020), this may be explained by the fact that at these sizes less than 0.3% of ByT5 parameters are locked in embedding tables and a larger amount of dense parameters is updated during training. mT5 parameters are instead dominated by the embedding tables, which are updated less often than the dense layers. In addition to that, ByT5-large is worse than ByT5-base at span labeling, which is a word level task. Both our observations confirm the findings in Xu et al. (2020).

[2] https://mmnlu-22.github.io

Model	IA	EM
ByT5-small	49.26	20.36
ByT5-base	64.3	33.47
ByT5-large	66.53	28.43
ByT5-xl	80.96	41.7
ByT5-xxl	81.73	38.28
mT5-small	51.75	17.59
mT5-base	55.91	17.73
mT5-large	67.23	25.14
mT5-xl	79.97	45.60
mT5-xxl	**82.44**	**50.21**

Table 1: Zero-shot *T5 parsers performance when training on English only.

Model	IA	EM
ByT5-small	85.59	66.60
ByT5-base	85.93	67.54
ByT5-large	84.02	62.92
ByT5-xl	87.01	68.29
ByT5-xxl	**87.27**	**68.66**
mT5-small	73.29	46.65
mT5-base	82.03	58.24
mT5-large	85.58	64.13
mT5-xl	87.24	68.47
mT5-xxl	86.79	63.33

Table 2: *T5 parsers performance when training on all the available gold data.

In the **synthetic data setting** (Table 3), IA almost matches the IA of models from the gold data setting. If we consider EM accuracy, we are only 5% points behind the upper bound performance of the multilingually supervised -xxl models (see Table 2). This indicates that synthetic data augmentation is a viable approach for the i18n of semantic parsers. Please refer to Table 9 in the appendix for results on individual languages.

4 Additional Experiments and Results

In zero-shot evaluations, English is the most studied language given the availability of labeled data. Recent work has shown that this language may not be the best at cross-lingual transfer (Turc et al., 2021). Since MASSIVE provides training and test data for all its languages, we can evaluate the zero-shot performance of each language. We train 51 ByT5-base model for a fixed number of steps

Model	IA	EM
ByT5-small	83.32	59.32
ByT5-base	84.59	61.24
ByT5-large	82.82	58.09
ByT5-xl	85.90	62.98
ByT5-xxl	86.48	**64.18**
mT5-small	73.64	43.19
mT5-base	80.79	51.76
mT5-large	83.99	57.43
mT5-xl	86.07	62.33
mT5-xxl	**86.69**	62.49

Table 3: *T5 parsers performance when training on English and synthetic TAF data.

(1k steps, 128 batch size) and collect the results on the development sets in Figure 2. By summing the EMs on rows we can understand how much a fine-tuning language (*donor*) improves the others. If we sum over columns, we can see how much transfer a target language (*receiver*) gets from the others. We report some statistics about best/worst donor/receiver languages in Table 4. Interestingly, English is not among the top donors, while it is the one that is being improved the most by other languages. We speculate that the better English LM representations may already have an intrinsic notion of semantic concepts that are then quickly individuated if supervision for such concepts is provided in other languages. From Figure 2, we see that some languages (am, sw, km, cy) clearly need annotated data. We hope that this map could help prioritize data collection efforts.

MASSIVE examples contain an interesting piece of metadata that indicates if an utterance has been translated and localized (i.e. original entities have been substituted with entities more culturally relevant for the target language), or translated only. We split the test sets in two parts according to this information and report in Figure 1 the EM accuracies of the same mT5-xxl model. We examine the three data settings studied in this paper. Accuracies on *localized* utterances are consistently lower. The performance difference in the synthetic data setting is relatively small but it still suggests that creating synthetic examples with entities that are *local* to the target language may improve the robustness of the parser.

In the appendix, we report the accuracy for each individual intent on the union of the test set ex-

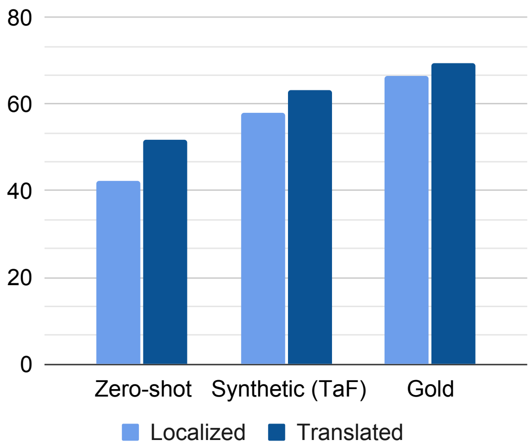

Figure 1: Differences in EM for an mT5-xxl model evaluated on queries of the test set that have been both translated and localized, vs only translated.

	Best to worst
Donor	fr, de, es, nl, pl, · · ·, mn, am, sw, km, cy
Receiver	en, de, pt, fr, sv, · · ·, zh, am, mn, sw, cy

Table 4: Top-5 Best/worst donor/receiver.

amples from all languages (Table 8). In Table 5, we report the 6 intents with the lowest accuracy. Most examples belong to the `GENERAL_QUIRKY` intent. The latter is likely a bucket intent covering all the utterances that are generic or out-of-domain (we could not find an exhaustive description of this intent in the SLURP dataset(Bastianelli et al., 2020)). The common parser mistake is to classify such queries as belonging to a more specific intent that can plausibly be associated with that query.

Finally, we compare our NMT translations of the training set with the corresponding gold translations produced by professional translators. We summarize the most interesting information in Table 6 (full comparison in Table 7 included in the appendix). Indic languages (*_IN and bn_BD) have an higher average match than other languages. This may suggest that translations in these languages are more unambiguous or that translators may have relied on a MT during the translation task.

Intent	**IA**	**Support**
`GENERAL_GREET`	19.6	51
`MUSIC_SETTINGS`	27.1	306
`AUDIO_VOLUME_OTHER`	54.9	306
`GENERAL_QUIRKY`	55.6	8619
`IOT_HUE_LIGHTON`	61.4	153
`MUSIC_DISLIKENESS`	74.5	204

Table 5: IA of the ByT5-xxl+TAF model for the lowest scoring intents (considering all languages).

Language sets	**Avg Match** (%)
All languages	21.3
All but Indic languages	17.3
Indic languages	50.8

Table 6: Percentages of NMT translations matching human translations in MASSIVE training set.

5 Related Work

Multilingual models are architecturally similar to monolingual transformer-based models but they are pretrained on multilingual corpora. These models include XLM (Lample and Conneau, 2019), XLM-R (Conneau et al., 2020) and mT5 (Xue et al., 2021), the multilingual version of T5 (Raffel et al., 2020). They all use a subword vocabulary, a choice that may result in poor performance for languages with limited amount of data (Wang et al., 2021). Token-free models such as ByT5 (Xu et al., 2020), Canine (Clark et al., 2022) and Charformer (Tay et al., 2022) were designed to avoid this issue and have been applied to popular multilingual benchmarks obtaining state-of-the-art results. In this work, we compare the multilinguality and the generative capabilities of mT5 and ByT5 in a massively multilingual semantic parsing task.

Data augmentation is the process of creating synthetic labeled data from available annotated examples. One approach in the multilinguality space is to translate annotated data in one language, e.g. English, to other languages. Neural machine translation is a strong baseline as it has been shown in recent cross-lingual evaluation benchmarks (Hu et al., 2020; Ladhak et al., 2020). While translation works quite well for classification tasks where the label is at instance level, sequence tagging or parsing tasks require an annotation projection step because labels are at token level. Translate-and-align methods use bilingual word aligners, statistical (Brown et al., 1993; Vogel et al., 1996; Och and Ney, 2000, 2003), and neural

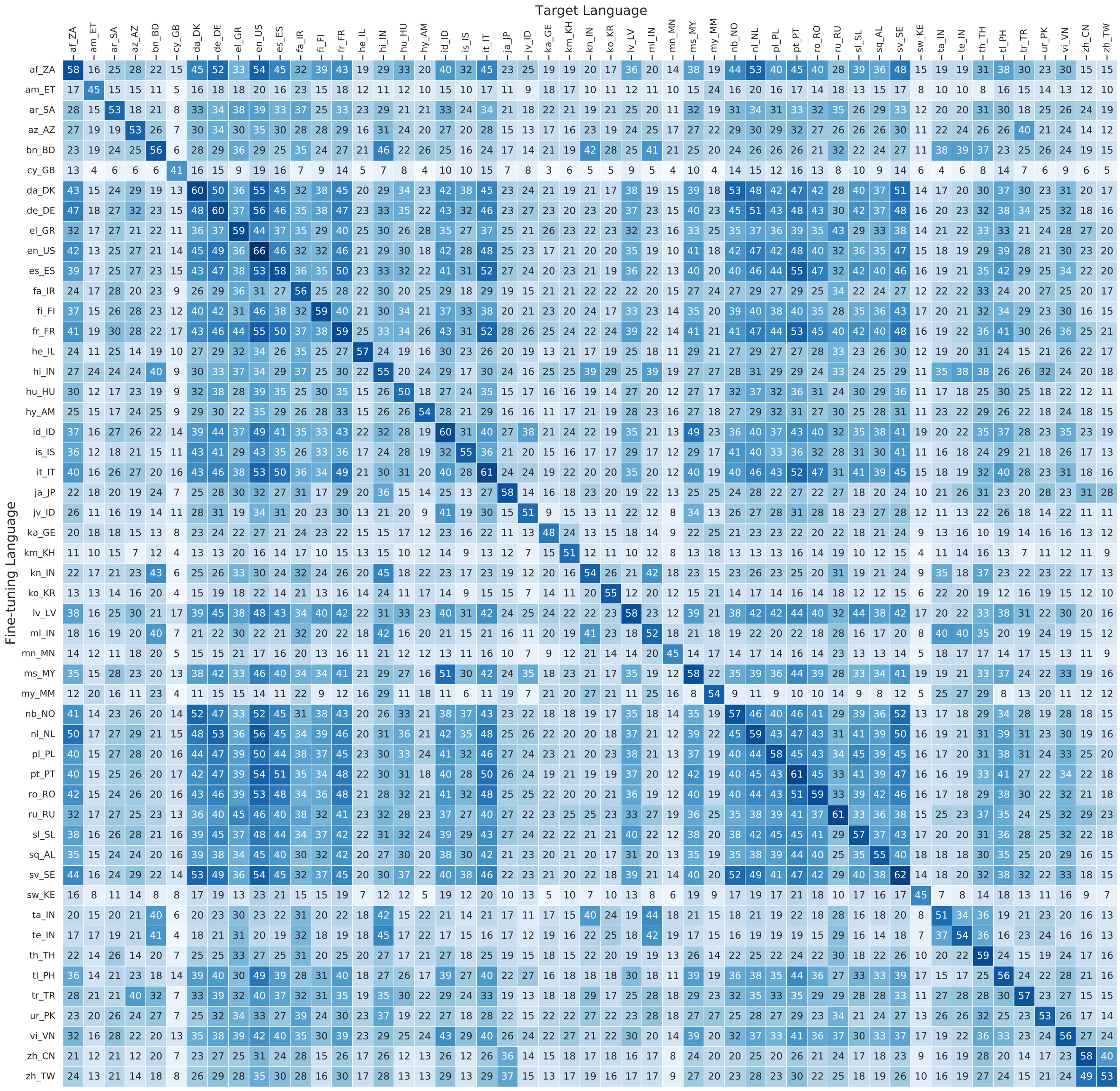

Figure 2: Zero-shot EM accuracies of individual ByT5-base models fine-tuned on a single language (y-axis) and evaluated on dev sets from all languages (x-axis).

(Schuster et al., 2019; Chen et al., 2020; Zenkel et al., 2020). More recent works removes this explicit alignment requirement (Dong and Lapata, 2018; Zhang et al., 2019; Wiseman et al., 2018). In our work, we use a label projection method based on pretrained language models (Nicosia et al., 2021) that reconstructs a full semantic parse from an utterance and a signature of the same parse.

6 Conclusions

In this paper, we evaluated ByT5 and mT5 (Xue et al., 2021) models in a massively multilingual semantic parsing task, showing that ByT5 is particularly competitive at smaller sizes. We have provided a map of the cross-lingual transfer for all the languages in MASSIVE and demonstrated that synthetic examples created with NMT are effective for building accurate semantic parsers.

Limitations

This work uses seq2seq models as parsers. Different output formats can yield better or worse results as shown in Paolini et al. (2021). We do not focus on tweaking formats or on modeling improvements such as constrained decoding for a more faithful generation. We adopt a compact output representation that reduces the text the model has to generate (and hallucinations) and gives us competitive results. In the cross-lingual transfer experiments, we train each model for a small fixed number of steps. If we train for longer, the representations start to change significantly and cross-lingual performances vary quite unpredictably. We leave for the future an investigation of the learning dynamics in this setting and the design of possible remedies.

References

Yoav Artzi and Luke Zettlemoyer. 2011. Bootstrapping semantic parsers from conversations. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 421–432.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262, Online. Association for Computational Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. 2020. Accurate word alignment induction from neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 566–576, Online. Association for Computational Linguistics.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

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.

Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2022. GlobalWoZ: Globalizing MultiWoZ to develop multilingual task-oriented dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1639–1657, Dublin, Ireland. Association for Computational Linguistics.

Li Dong and Mirella Lapata. 2018. Coarse-to-fine decoding for neural semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–742, Melbourne, Australia. Association for Computational Linguistics.

Kais Dukes. 2014. Semeval-2014 task 6: Supervised semantic parsing of robotic spatial commands. In *SemEval@ COLING*, pages 45–53.

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.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

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, Online. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021. Bitod: A bilingual multi-domain dataset for task-oriented dialogue modeling. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Olga Majewska, Evgeniia Razumovskaia, Edoardo Maria Ponti, Ivan Vulić, and Anna Korhonen. 2022. Cross-lingual dialogue dataset creation via outline-based generation.

Mehrad Moradshahi, Giovanni Campagna, Sina Semnani, Silei Xu, and Monica Lam. 2020. Localizing open-ontology QA semantic parsers in a day using machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5970–5983, Online. Association for Computational Linguistics.

Massimo Nicosia, Zhongdi Qu, and Yasemin Altun. 2021. Translate & Fill: Improving zero-shot multilingual semantic parsing with synthetic data. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3272–3284, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hong Kong. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, RISHITA ANUBHAI, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *International Conference on Learning Representations*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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. Association for Computational Linguistics.

Tom Sherborne, Yumo Xu, and Mirella Lapata. 2020. Bootstrapping a crosslingual semantic parser. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 499–517, Online. Association for Computational Linguistics.

Yi Tay, Vinh Q. Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2022. Charformer: Fast character transformers via gradient-based subword tokenization. In *International Conference on Learning Representations*.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *ArXiv*, abs/2106.16171.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2021. Multi-view subword regularization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 473–482, Online. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2018. Learning neural templates for text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3174–3187, Brussels, Belgium. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

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.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2020. End-to-end neural word alignment outperforms GIZA++. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1605–1617, Online. Association for Computational Linguistics.

Xiang Zhang, Shizhu He, Kang Liu, and Jun Zhao. 2019. AdaNSP: Uncertainty-driven adaptive decoding in neural semantic parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4265–4270, Florence, Italy. Association for Computational Linguistics.

A Comparing NMT with Gold Translations

In Table 7, we compare how many times the NMT translated utterances match the gold translations produced by professional translators. We restrict the match to utterances that have been translated and not localized in the target language, since NMT cannot perform the localization step. In addition, we preprocess all compared utterances with unicode normalization, we strip whitespaces and punctuation. In general, indic locales have higher match rates compared to other locales. Please also note that we translate English to pt_BR (Brazilian Portuguese) and this explains the low match for pt_PT.

B Intent Accuracy Performance

In Table 8, we report the accuracy for each individual intent on the union of the test set examples from all languages using ByT5-xxl + TAF.

C Performance on all Languages

In Table 9, we report Exact Match on all the 51 languages, for the three different experimental setups described in Section 3, across two models (mT5 and ByT5) and two model sizes (base and xxl).

Language	NMT vs Gold Translations (%)	Matches (#)	Non-localized sentences (#)
kn_IN	68.7	6524	9497
te_IN	54.1	4841	8941
bn_BD	52.6	4458	8471
ta_IN	48.3	4301	8898
hi_IN	46.5	4101	8827
nl_NL	38.5	3878	10 070
fr_FR	36.0	3736	10 385
ml_IN	34.7	2985	8607
tl_PH	34.0	3397	10 000
af_ZA	32.8	3160	9640
tr_TR	32.1	2998	9330
sw_KE	26.1	2336	8965
sv_SE	25.9	2465	9504
nb_NO	23.8	2402	10 083
vi_VN	21.6	2000	9255
ms_MY	21.6	1880	8702
jv_ID	21.1	1947	9208
pl_PL	21.0	2017	9618
da_DK	20.4	1933	9470
id_ID	20.4	1882	9227
es_ES	19.5	1876	9596
zh_CN	19.0	1661	8727
zh_TW	18.2	1638	8976
it_IT	17.9	1596	8916
fi_FI	17.5	1669	9558
ru_RU	17.4	1550	8912
hy_AM	16.9	1809	10 707
is_IS	16.1	1491	9270
km_KH	16.1	1491	9276
cy_GB	15.9	1578	9936
sl_SL	14.7	1313	8913
am_ET	14.6	1267	8658
hu_HU	14.5	1331	9198
ur_PK	14.4	1260	8761
de_DE	14.2	1422	9992
lv_LV	12.4	1071	8650
he_IL	12.3	1123	9159
sq_AL	12.2	1035	8460
az_AZ	12.1	1102	9081
th_TH	11.7	1041	8894
ro_RO	10.9	1001	9197
el_GR	10.5	934	8879
pt_PT	9.9	934	9392
ar_SA	9.9	871	8814
mn_MN	8.9	785	8826
fa_IR	8.3	718	8686
ja_JP	7.4	704	9487
ka_GE	7.4	701	9528
ko_KR	3.9	341	8804
my_MM	2.0	171	8765

Table 7: Number of verbatim matches between Gold translation and NMT translations.

Intent	IA	Support
GENERAL_GREET	19.6	51
MUSIC_SETTINGS	27.1	306
AUDIO_VOLUME_OTHER	54.9	306
GENERAL_QUIRKY	55.6	8619
IOT_HUE_LIGHTON	61.4	153
MUSIC_DISLIKENESS	74.5	204
DATETIME_CONVERT	75.6	765
IOT_WEMO_ON	76.3	510
PLAY_AUDIOBOOK	78.0	2091
TRANSPORT_QUERY	78.1	2601
RECOMMENDATION_EVENTS	78.3	2193
RECOMMENDATION_MOVIES	79.2	1020
CALENDAR_QUERY	80.6	6426
QA_FACTOID	82.4	7191
IOT_HUE_LIGHTUP	82.5	1377
LISTS_QUERY	82.6	2601
AUDIO_VOLUME_UP	83.0	663
SOCIAL_QUERY	83.9	1275
MUSIC_QUERY	84.0	1785
EMAIL_ADDCONTACT	84.5	612
MUSIC_LIKENESS	84.7	1836
EMAIL_QUERYCONTACT	84.8	1326
TAKEAWAY_QUERY	85.0	1785
LISTS_CREATEORADD	85.6	1989
QA_DEFINITION	86.3	2907
LISTS_REMOVE	86.3	2652
COOKING_RECIPE	86.6	3672
NEWS_QUERY	86.9	6324
PLAY_MUSIC	87.1	8976
TAKEAWAY_ORDER	87.3	1122
IOT_HUE_LIGHTDIM	87.4	1071
PLAY_PODCASTS	87.6	3213
PLAY_GAME	87.7	1785
ALARM_SET	89.5	2091
PLAY_RADIO	90.0	3672
CALENDAR_SET	90.2	10 659
RECOMMENDATION_LOCATIONS	90.4	1581
QA_MATHS	90.7	1275
AUDIO_VOLUME_DOWN	90.7	561
SOCIAL_POST	91.1	4131
IOT_WEMO_OFF	91.3	918
AUDIO_VOLUME_MUTE	91.7	1632
ALARM_QUERY	91.8	1734
GENERAL_JOKE	92.0	969
EMAIL_QUERY	93.0	6069
TRANSPORT_TICKET	93.1	1785
CALENDAR_REMOVE	93.4	3417
EMAIL_SENDEMAIL	94.0	5814
IOT_CLEANING	94.2	1326
WEATHER_QUERY	94.6	7956
IOT_HUE_LIGHTOFF	94.8	2193
TRANSPORT_TAXI	95.3	1173
IOT_HUE_LIGHTCHANGE	95.4	1836
ALARM_REMOVE	95.5	1071
QA_STOCK	95.6	1326
DATETIME_QUERY	95.8	4488
TRANSPORT_TRAFFIC	96.3	765
QA_CURRENCY	96.6	1989
IOT_COFFEE	97.9	1836

Table 8: IA of the ByT5-xxl+TAF model for all intents (all languages).

	Zero Shot				Synthetic (TAF)				Gold			
	base		xxl		base		xxl		base		xxl	
Language	mT5	ByT5	mT5	ByT5	mT5	ByT5	mT5	ByT5	mT5	ByT5	mT5	ByT5
af_ZA	21.6	**51.1**	58.0	**59.7**	53.7	**64.7**	65.6	**66.8**	59.4	**68.5**	65.9	**69.3**
am_ET	4.7	**15.9**	**40.7**	22.0	40.8	**54.4**	**61.2**	61.0	48.7	**61.3**	62.0	**65.8**
ar_SA	14.6	**27.8**	**43.6**	23.3	45.9	**56.1**	60.1	**60.5**	52.3	**64.7**	61.1	**66.0**
az_AZ	8.9	**31.2**	**41.8**	34.0	46.4	**61.6**	61.9	**63.6**	57.0	**69.0**	62.6	**69.6**
bn_BD	10.8	**19.5**	**45.9**	25.3	51.0	**62.1**	64.3	**65.6**	57.6	**67.6**	64.6	**69.5**
cy_GB	5.9	**16.4**	**42.8**	40.2	35.7	**56.1**	61.5	**64.2**	42.1	**65.3**	61.4	**69.2**
da_DK	30.2	**53.1**	**60.9**	54.2	57.8	**67.5**	67.3	**68.7**	64.4	**71.7**	67.9	**71.3**
de_DE	28.3	**55.3**	**59.8**	59.5	60.2	**67.8**	67.5	**68.8**	64.1	**70.4**	68.0	**70.2**
el_GR	17.4	**31.5**	**57.2**	27.9	55.5	**64.2**	65.5	**66.6**	62.0	**68.3**	66.6	**68.7**
en_US	65.5	**72.2**	**74.0**	73.3	68.5	**72.6**	**73.7**	73.0	68.9	**72.7**	**73.3**	72.6
es_ES	26.1	**50.8**	**55.6**	52.2	58.7	**65.1**	65.0	**65.9**	61.1	**67.2**	65.9	**66.2**
fa_IR	17.6	**32.8**	**54.4**	24.0	54.9	**62.2**	63.2	**64.4**	59.9	**69.1**	63.4	**69.7**
fi_FI	16.3	**36.9**	**52.5**	47.4	51.2	**65.9**	65.6	**68.2**	59.4	**71.1**	66.8	**71.5**
fr_FR	29.9	**53.5**	**58.5**	54.3	59.3	**64.4**	65.1	**65.6**	62.3	**66.5**	65.8	**67.2**
he_IL	9.7	**21.0**	**40.4**	24.0	50.1	**59.4**	61.0	**63.2**	57.5	**67.3**	62.3	**68.4**
hi_IN	14.1	**26.3**	**52.9**	26.2	54.4	**62.6**	64.2	**64.4**	59.3	**66.5**	64.5	**67.2**
hu_HU	17.5	**33.5**	**45.3**	32.9	51.8	**62.2**	64.2	64.2	58.2	**68.5**	65.2	**69.5**
hy_AM	11.7	**20.5**	**44.6**	24.7	49.8	**58.4**	60.3	**62.2**	57.8	**67.7**	61.7	**68.9**
id_ID	24.1	**48.3**	58.6	**61.5**	59.0	**64.6**	65.5	**67.1**	63.4	**68.8**	66.2	**69.0**
is_IS	11.6	**32.1**	**47.2**	31.7	47.6	**60.9**	63.4	**65.9**	54.6	**68.5**	63.4	**69.6**
it_IT	25.3	**52.5**	59.5	59.5	57.2	**63.0**	64.6	**65.5**	60.2	**67.6**	65.7	**67.3**
ja_JP	**26.8**	23.3	**46.6**	29.3	51.0	**55.6**	57.3	**58.8**	60.5	**65.8**	58.7	**67.0**
jv_ID	10.7	**22.9**	45.8	**46.2**	42.5	**58.9**	62.1	**63.9**	48.5	**66.5**	62.6	**68.5**
ka_GE	9.7	**17.9**	**39.9**	22.1	45.4	**52.9**	54.8	**57.1**	54.5	**63.8**	56.2	**66.8**
km_KH	11.4	**18.0**	**44.8**	23.6	39.2	**51.8**	51.7	**55.7**	54.7	**63.8**	54.3	**67.0**
kn_IN	8.8	**20.2**	**41.9**	25.4	47.4	**58.6**	55.8	**61.7**	52.1	**63.8**	56.6	**65.8**
ko_KR	11.0	**16.3**	**49.8**	24.8	54.1	**61.5**	65.6	**65.8**	60.2	**68.7**	66.4	**70.3**
lv_LV	11.6	**40.3**	**51.9**	33.7	52.4	**61.2**	63.0	**64.6**	59.0	**69.6**	64.1	**70.4**
ml_IN	10.1	**19.4**	**41.2**	25.8	47.9	**55.3**	55.0	**58.5**	59.4	**68.2**	55.6	**69.2**
mn_MN	7.4	**13.4**	**38.9**	22.2	46.9	**57.0**	60.2	**62.7**	53.8	**66.1**	61.5	**68.7**
ms_MY	21.7	**45.0**	54.8	**59.9**	57.1	**65.7**	67.7	**68.0**	60.6	**69.3**	68.4	**68.9**
my_MM	10.7	**13.8**	**48.7**	23.1	51.5	**59.8**	61.9	**66.1**	59.3	**68.8**	64.3	**72.6**
nb_NO	26.9	**50.6**	**60.7**	56.3	60.7	**68.0**	68.8	**70.2**	65.0	**70.5**	69.9	**70.7**
nl_NL	28.3	**55.2**	60.1	**63.3**	60.2	**66.5**	67.4	**67.5**	64.7	**68.4**	68.3	**70.0**
pl_PL	19.0	**47.1**	**50.7**	46.0	56.2	**61.8**	62.0	**63.3**	59.7	**65.9**	62.5	**66.5**
pt_PT	28.1	**52.0**	**60.8**	50.6	61.5	**65.9**	66.8	**67.6**	63.6	**68.7**	67.5	**68.2**
ro_RO	22.8	**45.7**	**57.4**	52.7	55.8	**64.5**	65.7	**67.1**	60.2	**68.5**	65.9	**69.6**
ru_RU	19.0	**26.1**	**49.0**	26.1	56.9	**61.6**	63.5	**63.8**	63.5	**68.8**	64.0	**69.5**
sl_SL	15.8	**43.7**	**52.8**	47.8	53.2	**63.5**	64.5	**64.8**	57.7	**68.0**	64.5	**68.8**
sq_AL	15.3	**42.1**	**48.0**	39.9	48.8	**61.1**	61.2	**63.5**	54.2	**68.9**	61.3	**68.5**
sv_SE	26.0	**54.4**	**61.8**	53.0	62.6	**70.1**	70.6	**71.1**	65.9	**72.0**	71.2	**71.5**
sw_KE	9.6	**15.6**	**44.0**	41.9	44.2	**58.7**	58.2	**59.6**	48.0	**66.3**	58.6	**66.8**
ta_IN	10.9	**19.9**	**41.1**	24.3	48.2	**55.5**	56.4	**58.3**	56.6	**64.9**	58.0	**66.0**
te_IN	7.8	**21.6**	**46.4**	25.1	43.6	**60.0**	55.4	**62.7**	51.4	**65.0**	55.1	**67.5**
th_TH	21.8	**31.3**	**55.0**	26.8	47.4	**62.1**	62.2	**66.9**	63.2	**72.0**	64.6	**74.2**
tl_PH	18.9	**42.0**	56.9	**58.7**	53.2	**62.4**	65.7	**66.1**	56.7	**66.5**	66.5	**68.5**
tr_TR	14.4	**35.2**	**48.4**	38.5	51.6	**64.9**	65.5	**66.2**	58.5	**69.4**	65.5	**69.4**
ur_PK	9.7	**22.7**	**49.2**	22.8	50.5	**59.5**	61.5	**61.9**	54.1	**63.3**	62.6	**65.7**
vi_VN	15.1	**35.1**	**55.9**	36.4	49.8	**57.5**	61.0	**62.3**	55.5	**67.0**	62.1	**68.2**
zh_CN	**22.1**	17.3	**31.7**	24.1	45.6	**54.1**	53.0	**57.9**	60.8	**65.9**	54.9	**66.6**
zh_TW	**21.2**	16.5	**32.4**	24.2	45.2	**51.8**	52.0	**54.5**	58.2	**62.2**	53.8	**63.9**
Average	17.7	**33.5**	**50.2**	38.3	51.8	**61.2**	62.5	**64.2**	58.2	**67.5**	63.3	**68.7**

Table 9: *T5 parsers Exact Match on individual languages in the Zero-Shot, TAF and Gold settings.