DAMP: Doubly Aligned Multilingual Parser for Task-Oriented Dialogue

Anonymous ACL submission

Abstract

Modern virtual assistants are powered by task-oriented dialogue systems with internal semantic parsing engines. In global markets such as India and Latin America, mixed language input from bilingual users is prevalent. Prior work has shown that multilingual transformer-based models exhibit worse multilingual transfer for semantic parsing than for other benchmark tasks. In this work, we improve zero-shot multilingual semantic parsing without harming supervised performance. First, we show that pretraining alignment objectives improve multilingual transfer while also reducing negative transfer to English. We then introduce a constrained optimization method to improve alignment using domain adversarial training. Our **D**oubly **A**ligned **M**ultilingual **P**arser (**DAMP**) improves mBERT transfer performance by 3x, 6x, and 81x on the Spanish-English Task Oriented Parsing, Hindi-English Task Oriented Parsing and Multilingual Task Oriented Parsing benchmarks respectively, and outperforms XLM-R and mT5-Large while using 3.2x fewer parameters.

1 Introduction

Task-oriented dialogue systems are the backbone of virtual assistants, one of the most direct and pervasive interactions between users and Natural Language Processing (NLP) technology. Semantic parsing converts unstructured text to structured representations grounded in task actions. Due to the conversational nature of the interaction between users and task-oriented dialogue systems, variation in speaker vocabulary, syntax, and register is especially pervasive. Such variation is an essential challenge for the inclusiveness and reach of virtual assistants which aim to serve a global and diverse userbase (Liu et al., 2021).

In this work, we are motivated by a common form of variation for bilingual speakers (Doğruöz et al., 2021): codeswitching. Codeswitching occurs in two forms which both affect task-oriented dialogue. Inter-sentential codeswitching appears through multilingual requests made by the same user during a dialogue:

Play all rap **music** on my iTunes
Toca toda la **música** rap en mi iTunes

Intra-sentential codeswitching appears through the user makes a single query using multiple languages:

Play toda la rap **music** en mi iTunes

Both forms are used by bilingual speakers (Dey and Fung, 2014) and undermine the reliability of location, primary language preference, and even language identification as a mechanism to route requests to an appropriate monolingual system (Barman et al., 2014). While zero-shot multilingual transfer is often used to reduce annotation costs, codeswitching makes it a key robustness feature.

However, zero-shot structured prediction and parsing is still a challenge for state-of-the-art multilingual models (Ruder et al., 2021), highlighting the need for improved methods beyond scale to achieve this goal. Fortunately, as a fundamental property of the task, these linguistically diverse inputs are grounded in a shared semantic output space. Each of the above outputs corresponds to:

[play_music:[genre:rap][platform:iTunes]**]**

The grounded nature of semantic parsing makes cross-lingual alignment natural for the task.

Figure 1 shows our successful pursuit of **double alignment** using both contrastive alignment pretraining and a novel constrained adversarial finetuning method. Our **D**oubly **A**ligned **M**ultilingual **P**arser (**DAMP**) achieves strong zero-shot performance on both multilingual (inter-sentential) and intra-sentential codeswitched data, making it a robust model for bilingual users without harming English performance. We contribute the following:

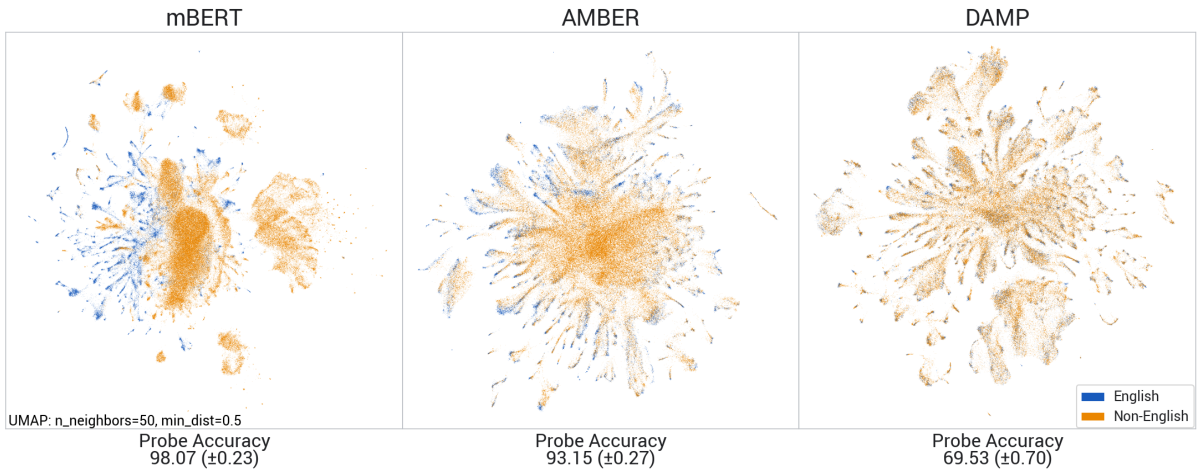

Figure 1: Language identification probe accuracy and visualizations of the embeddings from a multilingual transformer without alignment (mBERT), pretraining alignment alone (AMBER), and our proposed alignment regime of both contrastive pretraining and constrained adversarial finetuning (DAMP).

1. **Alignment Pretraining Effectiveness:** We first show multilingual BERT (mBERT) is ineffective for both categories of codeswitched data. We demonstrate that contrastive alignment pretraining with sentence-aligned monolingual data improves English, multilingual, and intra-sentential codeswitched semantic parsing performance.

2. **Constrained Adversarial Alignment:** We propose utilizing domain adversarial training to further improve alignment and transferability without labeled or aligned data. We introduce a novel constrained optimization method and demonstrate that it improves over prior domain adversarial training algorithms (Sherborne and Lapata, 2022) and regularization baselines (Li et al., 2018; Wu and Dredze, 2019) without hyperparameter tuning.

3. **Interpreting Alignment Improvements:** Through qualitative analysis, we find the improved parsing ability of DAMP is driven by a 6x improvement in prediction accuracy of the initial intent. We then provide evidence that our improvements are associated with measurable improvements in alignment. In Figure 1, we show improved alignment through embedding visualizations and a post-hoc linear probe on language prediction.

2 Related Work

Multilingual Language Model Alignment Massively multilingual transformers (MMTs) (Pires et al., 2019; Conneau et al., 2020a; Liu et al., 2020; Xue et al., 2021) have become the de-facto basis for multilingual NLP and are effective at intra-sentential codeswitching as well (Winata et al., 2021). These models appear to be effective at transfer as they implicitly perform alignment within representations of hidden states at later layers (Artetxe et al., 2020; Conneau et al., 2020b). Previously, many works have studied explicit objectives and training regimes to achieve stronger alignment, as representation alignment is an intuitively desirable property of transferable systems (Joulin et al., 2018; Artetxe et al., 2018; Artetxe and Schwenk, 2019).

Such models are remarkably robust for multilingual and intra-sentential codeswitching benchmarks (Aguilar et al., 2020; Hu et al., 2020; Ruder et al., 2021). However, the gap between performance on the training language and zero-shot targets is larger in task-oriented parsing benchmarks (Li et al., 2021; Agarwal et al.; Einolghozati et al., 2021), indicating weaker cross-lingual transfer efficiency likely caused by the language-specific structural knowledge needed for parsing.

Our work applies the pretraining regime from Hu et al. (2021), incorporating multiple explicit alignment objectives alongside traditional MMT pretraining. We show that this technique is effective both for semantic parsing, a new task, and intra-sentential codeswitching, a new linguistic domain.

Domain Adversarial Training The concept of using an adversary to regularize learning of undesirable features has been discovered and applied separately in transfer learning (Ganin et al., 2016), privacy preservation (Mirjalili et al., 2020), and algorithmic fairness (Zhang et al., 2018a). When applying this technique to transfer learning, Ganin et al. (2016) term this domain adversarial training. Due to its effectiveness in domain transfer learning, a variety of works have studied applications of domain adversarial learning to cross-lingual transfer (Guzman-Nateras et al., 2022; Lange et al., 2020; Joty et al., 2017). Most relevant, Sherborne and Lapata (2022) combines a multi-class language discriminator with translation loss to improve cross-lingual transfer.

We make the following 3 contributions to this space. Firstly, we show that token-level adversarial discrimination improves transfer to intra-sentential codeswitching without data of that form. Secondly, we show that binary discrimination is more effective than multi-class discrimination and provide intuitive reasoning for this surprising phenomenon. Finally, we remove the challenge of zero-shot hyperparameter search with a novel constrained optimization technique that can be configured a priori based on our alignment goals.

Preventing Multilingual Forgetting Beyond adversarial techniques, prior work has used regularization to maintain multilingual knowledge learned only during pretraining. Li et al. (2018) shows that penalizing distance from a pretrained model is a simple and effective technique to improve transfer. Using a much stronger inductive bias, Wu and Dredze (2019) freezes early layers of multilingual models to preserve multilingual knowledge. This leaves later layers unconstrained for task specific data. We are the first to compare such regularization to adversarial techniques and show that DAMP also improves over these techniques.

3 Methods

We utilize two separate stages of alignment to improve zero-shot transfer in DAMP. During pretraining, we propose to use contrastive learning to improve alignment amongst pretrained representations. During finetuning, we add **double** alignment through domain adversarial training using a binary language discriminator and a constrained optimization approach. We apply these improvements to the encoder of a pointer-generator network that copies and generates tags to produce a parse.

3.1 Baseline Architecture

Following Rongali et al. (2020), we use a pointer-generator network to generate semantic parses. We tokenize words $[w_0, w_1 \ldots, w_m]$ from the labeling scheme into sub-words $[s_{0,w_0}, \ldots, s_{n,w_0}, s_{0,w_1}, s_{n,w_m}]$ and retrieve hidden states $[\mathbf{h}_{0,w_0}, \ldots, \mathbf{h}_{n,w_0}, \mathbf{h}_{0,w_1} \ldots, \mathbf{h}_{n,w_m}]$ from our encoder. We use the hidden state of the first subword for each word to produce word-level hidden states:

$$[\mathbf{h}_{0,w_0}, \mathbf{h}_{0,w_1} \ldots, \mathbf{h}_{0,w_m}] \tag{1}$$

Using 1 as a prefix, we use a randomly initialized auto-regressive decoder to produce representations $[\mathbf{d}_0, \mathbf{d}_1 \ldots, \mathbf{d}_t]$. At each action-step a, we produce a generation logit vector using a perceptron to predict over the vocabulary of intents and slot types $\mathbf{g}_a$ and a copy logit vector for the arguments from the original query $\mathbf{c}_a$ using similarity with Eq. 1:

$$\mathbf{g}_a = MLP(\mathbf{d}_a) \tag{2}$$

$$\mathbf{c}_a = [\mathbf{d}_a^\top \mathbf{h}_{0,w_1}, \mathbf{d}_a^\top \mathbf{h}_{0,w_1}, \ldots \mathbf{d}_a^\top \mathbf{h}_{0,w_m}] \tag{3}$$

Finally, we produce a probability distribution $\mathbf{p}^a$ across both generation and copying by applying the softmax to the concatenation of our logits and optimize the negative log-likelihood of the correct prediction a':

$$\mathbf{p}^a = \sigma([\mathbf{g}_a; \mathbf{c}_a]) \tag{4}$$

$$L_s = -log(\mathbf{p}^a_{a'}) \tag{5}$$

3.2 Alignment Pretraining

We evaluate the contrastive pretraining process AMBER introduced by Hu et al. (2021) for semantic parsing. AMBER combines 3 explicit alignment objectives: translation language modeling, sentence alignment, and word alignment using attention symmetry. We hypothesize that this process of improving alignment in mBERT will be especially effective for semantic parsing due to the semantically aligned nature of the task and the importance of alignment for our randomly initialized decoder to perform on unseen languages.

Translation language modeling was originally proposed by Conneau and Lample (2019). This technique is a traditional masked language modeling task, but uses parallel sentences as input and masking tokens in each language. Since masked words can be unmasked in the parallel sentences, this encourages the model to align word and phrase level representations so that they can be used interchangeably across languages.

Sentence alignment (Conneau et al., 2018) directly optimizes the similarity of representations across languages using a siamese network training process. Given a batch of English sentences and their translations, the model is trained to predict the correct translation for a mention with respect to in-batch negative translations. For pooled representation $\mathbf{e}_i$ of each English sentence with a batch of possible translations B including true translation t', the loss is computed by producing a logit vector using the inner product, normalizing using the softmax function, and computing negative log-likelihood:

$$L(\mathbf{e}_i, \mathbf{t}', N)_{sa} = \log\left(\frac{\mathbf{e}_i^\top \mathbf{t}'}{\sum_{t_i \in B} \mathbf{e}_i^\top \mathbf{t}_i}\right) \quad (6)$$

Finally, AMBER uses a loss function from Cohn et al. (2016) which encourages word level alignment by optimizing the symmetry of attention across languages. For attention head $h \in H$, a sentence in language S, and its translation in language T, we compute an attention matrix $A^h_{S\to T} \in \mathbb{R}^{M\times N}$ from S to the translation and the attention matrix of the translation to S. $A^h_{T\to S} \in \mathbb{R}^{N\times M}$. The loss is then computed as the average trace similarity between these matrices for all heads H:

$$L(S, T) = 1 - \frac{1}{H}\sum_{h\in H}\frac{\mathrm{tr}(A^h_{S\to T}{}^\top A^h_{T\to S})}{\min(M, N)} \quad (7)$$

3.3 Adversarial Alignment

We build on the domain adversarial training process of Ganin et al. (2016). First, we use a token-level language discriminator to get aligned representations at the word level. Unlike prior work, we propose and justify a binary scheme that classifies tokens as English or Non-English rather than the standard multi-class language discriminator. Finally, we introduce a general constrained optimization approach for domain adversarial training and apply it to cross-lingual alignment.

Token-Level Discriminator Similar to Ganin et al. (2016), we train a discriminator to distinguish between in-domain training data and unlabeled out-of-domain data. Our method assumes access to labeled training queries in one language, in this case English, and unlabeled multilingual queries which target the same intents and slots. Data from

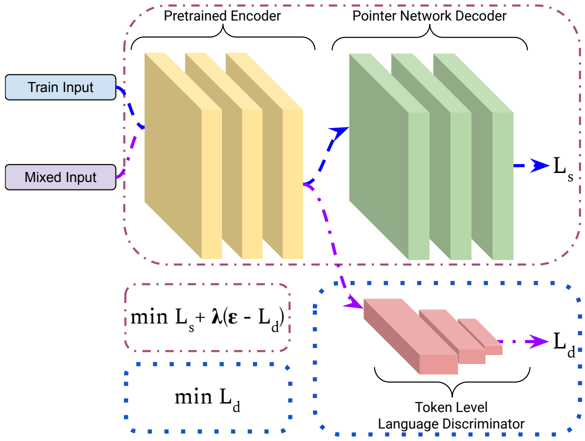

Figure 2: An overview of the adversarial alignment procedure. An adversarial model distinguishes English and Non-English examples with L_d. With $L_d \geq \epsilon$ as a constraint, the generator optimizes the Lagrangian dual.

each language is shuffled together with even sampling to create a dataset with equal amounts of each language.

We use a multilayer perceptron to predict the probability $p = P(E|h_{0,w_n})$ that a token with true label y is English or Non-English given hidden representations from Eq. 1. Our discriminator loss is traditional binary cross-entropy loss:

$$L_d = -(y\log(p) + (1-y)\log(1-p)) \quad (8)$$

This varies from prior work using domain adversarial training for multilingual robustness (Lange et al., 2020; Sherborne and Lapata, 2022) which performs multi-class classification across all languages and uses the negative log-likelihood of the correct class as the loss function. While this loss function is intuitively correct for the discriminator, it allows the generator to optimize towards maxima which do not benefit multilingual transfer.

First, suppose we have labeled data in English and unlabeled data in Spanish and French. The goal of the multi-class adversary is to predict English, Spanish, or French for each token while the encoder is to minimize the ability of the adversary to recover the correct language.

Even before adversarial training, the adversary is likely to struggle with tokens that are already well aligned across languages. For example, "dormir" in the Spanish sentence "recuérdame ir a dormir temprano (remind me to go to sleep early)" will be well aligned between French and Spanish since "dormir" translates to "to sleep" in both languages.

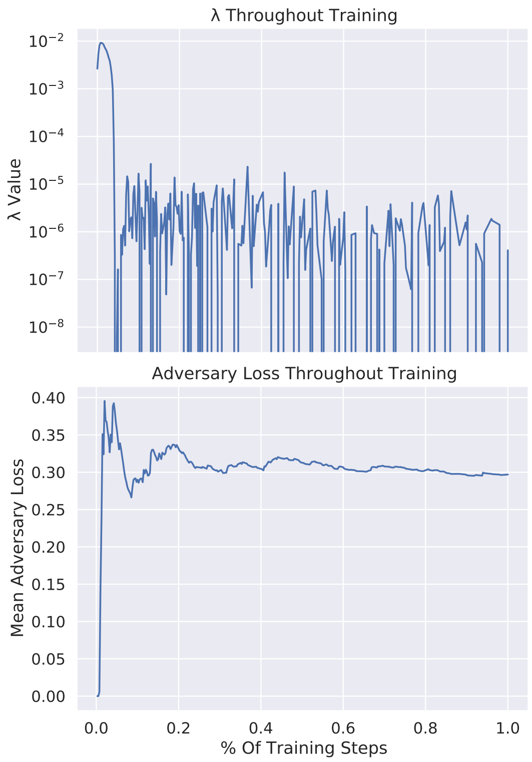

Figure 3: The top plot shows the learned schedule for the weight λ. The bottom plot shows the adversarial loss which converges to our constraint using this λ schedule.

This means the encoder can simply maintain alignment for the token "dormir" across French and Spanish, making it impossible for the adversary to recover the correct language. Doing so maximizes the multi-class adversarial loss but does not improve alignment between "dormir" and the English "to sleep" in our labeled data. In this extreme example, we highlight that multi-class alignment can be maximized without improving transferability from English at all.

Using a binary classifier removes such in-optimal solutions. Since each token is classified purely as English or Non-English, all tokens are aligned to an English equivalent. This prevents alignment between other non-English languages from leading to poor alignment with English.

Constrained Optimization Traditionally, domain adversarial training uses a gradient reversal layer (Ganin et al., 2016) to allow the generator to maximize adversary loss L_d weighted by hyperparameter λ while minimizing task loss L_s. For the generator, this is effectively equivalent to optimizing a linear combination of the terms:

$$L = L_s - \lambda L_d \tag{9}$$

However, selecting a schedule for λ presents a challenge in the zero-shot setting. Since the reverse validation procedure used to select the λ schedule by Ganin et al. (2016) assumes only one target domain, multilingual works such as Sherborne and Lapata (2022) opt to simply perform a linear search using the in-domain development set s. While simple, this approach ignores transfer performance entirely when weighing adversary loss. To address this, we propose a novel method of weighing adversarial loss using constrained optimization.

If our token representations are exactly aligned across languages, they are indistinguishable by any classifier. A well-suited adversary in this case will predict English and Non-English with $P = 0.5$ since it cannot perform better than chance. Such a model receives a loss of 0.3 for all inputs. Achieving a larger adversary loss is impossible in equilibrium since the adversary can decrease loss by predicting $P = 0.5$ regardless of the ground truth labels.

This reasoning provides a clear constraint on our desired adversarial loss. In alignment, the L_d should be no less than 0.3, which we call ϵ. We optimize the task loss L_s according to this constraint using back-propagation alone with the differential method of multipliers (Platt and Barr, 1987). The differential method of multipliers first relaxes the constrained problem to its Lagrangian dual:

$$L = L_s + \lambda(\epsilon - L_d) \tag{10}$$

λ is treated as a learnable parameter and optimized by stochastic gradient descent to maximize the value of $\lambda(\epsilon - L_d)$. In plain terms, this causes the value of λ to increase when $\epsilon > L_d$ and decrease when $\epsilon < L_d$. This learns a schedule for λ which weights the adversarial penalty according to its performance. We show the learned schedule of lambda in Figure 3 and demonstrate it causes the adversary loss term to converge to our constraint $\epsilon = 0.3$.

4 Experiments

We evaluate the effects of our techniques on three benchmarks for task-oriented semantic parsing with hierarchical parse structures. Two of these datasets evaluate robustness to intra-sentential codeswitching (Einolghozati et al., 2021; Agarwal et al.) and the third uses multilingual data to evaluate robustness to inter-sentential codeswitching (Li et al., 2021). Examples are divided as originally released into training, evaluation, and test data at a ratio of 70/10/20. We consider the limitations of these experiments in Appendix A.

4.1 Datasets

Multilingual Task Oriented Parsing (MTOP) Li et al. (2021) introduced this benchmark to evaluate multilingual transfer for a difficult compositional parse structure. The benchmark contains queries in English, French, Spanish, German, Hindi, and Thai. Zero-shot performance on this benchmark to evaluates inter-sentential codeswitching robustness. Each language has approximately 15,000 total queries which cover 11 domains with 117 intents and 78 slot types.

Hindi-English Task Oriented Parsing (CST5) Agarwal et al. construct a benchmark of Hindi-English intra-sentential codeswitching data using the same label space as the second version of the English Task Oriented Parsing benchmark (Chen et al., 2020). As part of preprocessing, we use Zhang et al. (2018b) to identify and transliterate Romanized Hindi tokens to Devanagari. There are 125,000 in English and 10,896 queries in Hindi-English which cover 8 domains with 75 Intents and 69 Slot Types.

Codeswitching Task Oriented Parsing (CSTOP) Einolghozati et al. (2021) is a benchmark of Spanish-English codeswitching data. While the dataset was released with a corresponding English dataset in the same label space, that data is now unavailable. Therefore, we construct an artificial dataset in the same label space using Google Translate on each segment of the structured Spanish-English training data. While the resulting English data is noisy, it provides an estimate of zero-shot transfer from English to Spanish-English codeswitching. The resulting dataset has 5,803 queries in both English and Spanish-English which cover 2 domains with 19 Intents and 10 Slot Types.

4.2 Results

For all benchmarks, we use the respective English data for training and development sets for early stopping. We report Exact Match (EM) accuracy on the English test split and zero-shot results on all other test splits. In all tables, bold results using are marked as significant ($p = 0.05$) using the bootstrap confidence interval for one run † (Dror et al., 2018).

We use the same hyperparameter configurations for all settings. The encoder uses the mBERT architecture (Pires et al., 2019). The decoder is a randomly initialized 4-layer, 8-head vanilla transformer for comparison with the 4-layer decoder structure used in Li et al. (2021). We use AdamW and optimize for 1.2 million training steps using a learning rate of $2e{-5}$, batch size of 16, and decay the learning rate to 0 throughout of training. We train on a Cloud TPU v3 Pod for approximately 4 hours for each dataset. For all adversarial experiments, we use the unlabeled queries from MTOP as training data for our discriminator and select a loss constraint ϵ of 0.3 as justified in 3.3.

MTOP In Table 1, we report the results of our architecture with mBERT, AMBER, and DAMP compared to existing baselines from prior work: XLM-R with a pointer-generator network (Li et al., 2021) and a finetuned MT5 (Nicosia et al., 2021).

Despite being a strong baseline for other tasks (Wu and Dredze, 2019; Aguilar et al., 2020; Liang et al., 2020; Hu et al., 2020; Ruder et al., 2021), mBERT alone is ineffective at cross-lingual transfer for compositional semantic parsing achieving an average multilingual accuracy of 0.5.

The AMBER pretraining process significantly improves accuracy for all languages to an average of 23.6. Average accuracy across the 5 Non-English languages improves by 47x. English accuracy also improves to 84.2 from 78.6, instead of suffering negative transfer (Wang et al., 2020).

DAMP further improves accuracy over AMBER by 1.8x to 42.2, outperforming both mT5-Large (31.4) and XLM-R (38.8). mT5-XXL maintains state-of-the-art performance of 55.1 but requires 33x more parameters and multiple GPUs for inference which heavily limits use.

Accuracy for each language is improved by at least 10 points, with Hindi and Thai, the most distant testing languages from English, having the largest improvements of +20.7 and +26.5 respectively. DAMP improves over the mBERT baseline by 84x without architecture changes or additional inference cost.

CST5 & CSTOP In Table 3, we report the results on both intra-sentential codeswitching benchmarks. For Hindi-English, we compare the MT5-small and MT5-XXL baselines from Agarwal et al..

AMBER again leads to a performance improvement for both CST5 and CSTOP, across English

	en	es	fr	de	hi	th	Avg(5 langs)	Parameters	Ratio
XLM-R	83.9	50.3	43.9	**42.3**	**30.9**†	26.7	38.8	550M	3.2x
mT5-Large	83.2	40.0	41.1	36.2	16.5	23.0	31.4	550M	3.2x
mT5-XXL	86.7	62.4	63.7	57.1	43.3	49.2	55.1	6.5B	33x
mBERT	78.6	0.5	1.0	0.9	0.1	0.1	0.5	172M	1x
AMBER	**84.2**	46.4	35.8	26.3	6.7	2.7	23.6	172M	1x
DAMP	83.5	**56.8**†	**55.6**†	42.2	27.4	**29.2**†	**42.2**†	172M	1x

Table 1: Exact Match (EM) accuracy scores on the MTOP dataset. XLM-R and mT5 results from Li et al. (2021) and Nicosia et al. (2021) respectively. Best results for models which fit on a single consumer GPU in bold.

	CST5		CSTOP		
	en	hi-en	en	es-en	Ratio
mT5-Small	-	6.4	-	-	0.9x
mT5-XXL	-	20.3	-	-	33x
mBERT	84.4	3.8	81.2	27.7	1x
AMBER	**85.8**	16.7	**86.7**†	79.3	1x
DAMP	85.6	**20.5**†	86.0	**80.3**†	1x

Table 2: Exact Match (EM) accuracy scores for intra-sentential codeswitching benchmarks CST5 and CSTOP. mT5 results from Agarwal et al.. Best results in bold.

(+1.4, +5.5) and codeswitched (+12.9, +52.4) data. DAMP also further improves transfer results (+3.8, +1.0) at the cost of small losses in English performance (-0.2, -0.7). DAMP achieves a new state-of-the-art of 20.5 on zero-shot transfer for CST5, outperforming even MT5-XXL (20.3). Since both alignment stages have word-level objectives, we hypothesize that the word-level inductive bias provides benefits for intra-sentential codeswitching despite lacking explicit codeswitching supervision.

4.3 Adversary Ablation

In Table 4.3, we isolate the effects of our contributions to domain adversarial training with an ablation study. While all adversarial variants improve transfer results, the usage of a binary adversary and our constrained optimization technique improve adversarial results independently and in combination. Notably, the multi-class adversary without constrained optimization is equivalent to (Sherborne and Lapata, 2022) using AMBER. DAMP improves over this prior adversarial technique by 9.9, 6.4, and 0.9 EM accuracy points on MTOP, CST5, and CSTOP respectively.

We also compare adversarial training to regularization techniques used in cross-lingual learning. We experiment with freezing the first 8 layers of the encoder (Wu and Dredze, 2019) and using the L_1

	MTOP		CST5		CSTOP	
	en	Avg	en	hi-en	en	es-en
Alignment Ablation						
mBERT	78.6	0.5	84.4	3.7	81.2	27.7
AMBER	**84.2**	23.6	**85.8**	16.7	**86.7**	79.3
+ Multi	84.0	32.3	85.5	14.1	85.0	79.4
+ Constr.	82.7	33.7	85.6	13.8	85.1	**80.3**
+ Binary	83.8	35.8	**85.8**	18.4	86.3	78.1
+ Constr.	83.5	**42.2**†	85.6	20.5	86.0	**80.3**
Regularization Baselines						
+ Freeze	82.6	32.0	85.2	**24.6**†	85.5	77.2
+ L_2 Norm	81.3	35.5	81.6	22.5	83.4	77.5
+ L_1 Norm	78.6	36.4	80.7	18.7	81.1	69.8

Table 3: Exact Match (EM) accuracy scores for multi-class and binary discriminators with and without our constrained optimization technique and regularization.

	en	es	fr	de	hi	th	Avg
mBERT	94.7	15.3	17.0	10.7	7.0	8.2	11.6
AMBER	**96.4**	78.7	71.3	66.3	32.5	26.5	55.1
DAMP	**96.4**	**89.0**†	**86.4**†	**80.5**†	**76.6**†	**74.4**†	**81.4**†

Table 4: Intent Prediction accuracy for each language on the MTOP dataset for mBERT, AMBER, and DAMP.

and L_2 norm penalty (Li et al., 2018). Adversarial learning outperforms these baselines on MTOP and CSTOP while model freezing and L_2 norm penalization outperform adversarial learning on CST5. However, adversarial learning is the only method that improves across all benchmarks.

4.4 Improvement Analysis

Since exact match accuracy is a strict metric, we analyze our improvements through qualitative analysis. We filtered to examples that DAMP predicts correctly but AMBER and mBERT do not. We then randomly sampled 20 examples from each language for manual evaluation.

We noted that improvements in intent prediction led to a large portion of the gain. If intent prediction fails, the rest of the auto-regressive decoding goes awry as the decoder attempts to generate valid slot types for that intent. We report intent prediction results across the test dataset in Table 4.

In general, these improvements follow a trend from nonsensical errors to reasonable errors to correct. For example, given the French phrase "S'il te plait appelle Adam." meaning "Please call Adam."", mBERT predicts the intent *QUESTION_MUSIC*, AMBER predicts *GET_INFO_CONTACT*, and DAMP predicts the correct *CREATE_CALL*.

Within the slots themselves, the primary improvements noted in DAMP are of less clear practical importance. DAMP more consistently abides by the annotation guideline preference of not including articles and prepositions such as "du", "a", "el", and "la" inside the slot boundaries.

We present the full sample of examples used for this analysis in Tables 5-9 in the Appendix.

5 Alignment Analysis

We analyze to what degree each method achieves our desired alignment goals beyond empirical effectiveness using two methods in Figure 1. First, we use a two-dimensional projection of the resulting encoder embeddings to provide a visual intuition for alignment. Then, we quantitatively evaluate alignment using a post-hoc linear probe.

5.1 Embedding Space Visualization

We visualize the embedding spaces of each model variant on each MTOP test set using Universal Manifold Approximation and Projection (UMAP) (McInnes et al., 2018). Our visualization of mBERT provides a strong intuition for its poor results, as English and Non-English data form linearly separate clusters even within this reduced embedding space. By using AMBER instead, this global clustering behavior is removed and replaced by small local clusters of English and Non-English data. Finally, DAMP produces an embedding space with no clear visual clusters of Non-English data without English data intermingled.

5.2 Post-Hoc Probing

We evaluate improvements to alignment quantitatively. While prior work used the performance of the training adversary to confirm alignment (Sherborne and Lapata, 2022), other studies have shown that new discriminators trained on the final model can recover information that the original adversary could not (Elazar and Goldberg, 2018; Ravfogel et al., 2022). Therefore, we train a post-hoc linear probe on representations generated by frozen versions of each of our model variants after training using 10-fold cross-validation.

Supporting the visual intuition, probe performance decreases with each stage of alignment. On mBERT, the discriminator achieves 98.07 percent accuracy indicating poor alignment. While AMBER decreases discriminator performance, it still achieves 93.15 percent accuracy indicating the need for further removal. Finally, DAMP results in a 23.62 point drop in discriminator accuracy to 69.53. However, the post-hoc adversary accuracy is still far above chance despite our training adversary converging to close-to-random accuracy. This indicates the possibility of further alignment improvements.

6 Conclusions and Future Work

In this work, we introduce the Doubly Aligned Multilingual Parser (DAMP), a semantic parsing training regime that uses explicit alignment objectives in pretraining and finetuning.

We first illustrated that contrastive alignment objectives in pretraining significantly improve zero-shot semantic parsing performance across multilingual and intra-sentential codeswitching data. However, this treatment alone leaves many poorly aligned clusters after finetuning. We therefore contribute a novel constrained optimization technique for domain adversarial training and apply it to language alignment using a binary classifier. Empirically, we show that for both multilingual and intra-sentential codeswitching DAMP improves over the mBERT baseline by large margins and outperforms larger models.

We identify 2 future application areas for our **D**oubly **A**ligned **M**ultilingual **P**arser (**DAMP**):

- Our novel constrained domain adversarial training is not specific to cross-lingual transfer. We encourage evaluations of this technique for domain adversarial training more broadly.
- We evaluate DAMP on multiple task-oriented semantic parsing benchmarks, however DAMP is amenable to any multilingual parsing task. We encourage the usage of DAMP to improve other parsing tasks

References

Anmol Agarwal, Jigar Gupta, Rahul Goel, Shyam Upadhyay, Pankaj Joshi, and Rengarajan Aravamudhan. CST5: Data augmentation for code-switched semantic parsing. In-Review.

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23, Doha, Qatar. Association for Computational Linguistics.

Xilun Chen, Asish Ghoshal, Yashar Mehdad, Luke Zettlemoyer, and Sonal Gupta. 2020. Low-resource domain adaptation for compositional task-oriented semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5090–5100, Online. Association for Computational Linguistics.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 876–885, San Diego, California. 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. 2020a. 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*, volume 32. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Anik Dey and Pascale Fung. 2014. A Hindi-English code-switching corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. 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 1654–1666, Online. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Arash Einolghozati, Abhinav Arora, Lorena Sainz-Maza Lecanda, Anuj Kumar, and Sonal Gupta. 2021. El volumen louder por favor: Code-switching in task-oriented semantic parsing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1009–1021, Online. Association for Computational Linguistics.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages

11–21, Brussels, Belgium. Association for Computational Linguistics.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030.

Luis Guzman-Nateras, Minh Van Nguyen, and Thien Nguyen. 2022. Cross-lingual event detection via optimized adversarial training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5588–5599, Seattle, United States. Association for Computational Linguistics.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3633–3643, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *ICML*, pages 4411–4421.

Shafiq Joty, Preslav Nakov, Lluís Màrquez, and Israa Jaradat. 2017. Cross-language learning with adversarial neural networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 226–237, Vancouver, Canada. Association for Computational Linguistics.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Lukas Lange, Anastasiia Iurshina, Heike Adel, and Jannik Strötgen. 2020. Adversarial alignment of multilingual models for extracting temporal expressions from text. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 103–109, Online. Association for Computational Linguistics.

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.

Xuhong Li, Yves Grandvalet, and Franck Davoine. 2018. Explicit inductive bias for transfer learning with convolutional networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2825–2834. PMLR.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Jiexi Liu, Ryuichi Takanobu, Jiaxin Wen, Dazhen Wan, Hongguang Li, Weiran Nie, Cheng Li, Wei Peng, and Minlie Huang. 2021. Robustness testing of language understanding in task-oriented dialog. 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 2467–2480, Online. Association for Computational Linguistics.

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

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Vahid Mirjalili, Sebastian Raschka, and Arun Ross. 2020. Privacynet: semi-adversarial networks for multi-attribute face privacy. *IEEE Transactions on Image Processing*, 29:9400–9412.

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.

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, Florence, Italy. Association for Computational Linguistics.

John Platt and Alan Barr. 1987. Constrained differential optimization. In *Neural Information Processing Systems*.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. 2022. Linear adversarial concept erasure. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18400–18421. PMLR.

Subendhu Rongali, Luca Soldaini, Emilio Monti, and Wael Hamza. 2020. Don't parse, generate! a sequence to sequence architecture for task-oriented semantic parsing. In *Proceedings of The Web Conference 2020*, pages 2962–2968.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Sherborne and Mirella Lapata. 2022. Zero-shot cross-lingual semantic parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4134–4153, Dublin, Ireland. Association for Computational Linguistics.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. 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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018a. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA. Association for Computing Machinery.

Yuan Zhang, Jason Riesa, Daniel Gillick, Anton Bakalov, Jason Baldridge, and David Weiss. 2018b. A fast, compact, accurate model for language identification of codemixed text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 328–337, Brussels, Belgium. Association for Computational Linguistics.

A Limitations

This work only carries out experiments using English as the base training language for domain adversarial transfer. It is possible that domain adversarial transfer has a variable effect depending on the training language from which labeled data is used. Additionally, while typologically diverse, all but one language used in our evaluation is of Indo-European origin.

	Input	Model	Outputs
Spanish	¿ dónde trabaja packy ahora ?	mBERT AMBER DAMP	[in:get_contact [sl:contact trabaja ahora]] [in:get_location [sl:contact packy]] [in:get_employer [sl:contact packy]]
	infórmame sobre la lluvia .	mBERT AMBER DAMP	[in:send_message [sl:recipient sobre] [sl:content_exact la lluvia]] [in:get_weather [sl:weather_attribute infórmame sobre la lluvia]] [in:get_weather [sl:weather_attribute lluvia]]
	enumerar mis alarmas .	mBERT AMBER DAMP	[in:create_alarm [sl:alarm_name enumerar mis]] [in:create_alarm] [in:get_alarm]
	llama a andy	mBERT AMBER DAMP	[in:send_message [sl:recipient llama andy]] [in:create_call [sl:contact a andy]] [in:create_call [sl:contact andy]]
	agrega a kelly a la llamada .	mBERT AMBER DAMP	[in:play_music [sl:music_artist_name agrega kelly] [sl:music_track_title la llamada]] [in:play_music [sl:music_artist_name agrega kelly] [sl:music_track_title la llamada]] [in:update_call [sl:contact_added kelly]]
	pausar y apagar llamada	mBERT AMBER DAMP	[in:get_recipes [sl:recipes_dish pausar] [sl:recipes_source llamada]] [in:end_call] [in:switch_call]
	¿ necesito un gran abrigo ?	mBERT AMBER DAMP	[in:get_education_time [sl:contact necesito] [sl:location gran] [sl:type_relation abrigo]] [in:get_info_recipes [sl:recipes_dish gran abrigo]] [in:get_weather [sl:weather_attribute abrigo]]
	llámame al mediodía	mBERT AMBER DAMP	[in:create_call [sl:contact al]] [in:create_call [sl:contact llámame al mediodía]] [in:create_alarm [sl:date_time al mediodía]]
	reproduce 1470 en la radio	mBERT AMBER DAMP	[in:is_true_recipes [sl:recipes_meal reproduce la] [sl:music_type radio]] [in:replay_music [sl:music_radio_id 1470] [sl:music_type radio]] [in:play_music [sl:music_radio_id 1470] [sl:music_type radio]]
	¿ cómo va el temporizador ?	mBERT AMBER DAMP	[in:get_contact [sl:contact_related va temporizador]] [in:get_reminder [sl:amount cómo temporizador]] [in:get_timer [sl:method_timer temporizador]]
	¿ se pronostican tormentas ?	mBERT AMBER DAMP	[in:get_education_degree [sl:contact se pronostican] [sl:contact_related tormentas]] [in:get_info_recipes [sl:recipes_unit_nutrition pronostican tormentas]] [in:get_weather [sl:weather_attribute tormentas]]
	¿ cómo hago un roux ?	mBERT AMBER DAMP	[in:get_contact [sl:contact cómo hago] [sl:location roux]] [in:get_info_recipes] [in:get_recipes [sl:recipes_dish roux]]
	dame el tiempo en australia	mBERT AMBER DAMP	[in:question_news [sl:news_topic el australia]] [in:get_event [sl:location australia]] [in:get_weather [sl:location australia]]
	muéstra me gente libre	mBERT AMBER DAMP	[in:get_recipes [sl:recipes_dish gente libre]] [in:send_message [sl:recipes_included_ingredient muéstra] [sl:recipes_rating gente libre]] [in:get_availability]
	¿ habrá granizo ?	mBERT AMBER DAMP	[in:get_contact [sl:type_relation granizo]] [in:get_weather [sl:contact granizo]] [in:get_weather [sl:weather_attribute granizo]]
	tiempo en nueva york	mBERT AMBER DAMP	[sl:music_genre en york]] [in:add_time_timer [sl:method_timer tiempo] [sl:location nueva york]] [in:get_weather [sl:location nueva york]]
	¿ quién fue a yale ?	mBERT AMBER DAMP	[in:get_contact [sl:type_relation yale]] [in:get_info_contact [sl:contact yale]] [in:get_contact [sl:school yale]]
	ponme en línea .	mBERT AMBER DAMP	[in:play_music [sl:music_track_title ponme en]] [in:end_call] [in:set_available]
	haz una llamada a mi papá	mBERT AMBER DAMP	[in:get_recipes [sl:recipes_dish haz mi]] [in:create_call [sl:contact [in:get_contact [sl:contact_related mi papá]]]] [in:create_call [sl:contact [in:get_contact [sl:contact_related mi] [sl:type_relation papá]]]]
	¿ cuándo comienza a llover ?	mBERT AMBER DAMP	[in:get_contact [sl:type_relation comienza]] [in:get_details_news] [in:get_weather [sl:weather_attribute llover]]

Table 5: Full Table of 100 Sampled Spanish Results from Qualitative Analysis.

	Input	Model	Outputs
French	prends lauren au téléphone	mBERT AMBER DAMP	[in:get_alarm [sl:ordinal prends lauren]] [in:update_call [sl:contact_added lauren]] [in:create_call [sl:contact lauren]]
	joue du frank ocean .	mBERT AMBER DAMP	[in:like_music [sl:music_provider_name frank ocean]] [in:play_music [sl:music_artist_name du ocean]] [in:play_music [sl:music_artist_name frank ocean]]
	comment faire un roux ?	mBERT AMBER DAMP	[in:get_weather [sl:location comment faire] [sl:location roux]] [in:get_info_recipes] [in:get_recipes [sl:recipes_dish roux]]
	nouveau rappel .	mBERT AMBER DAMP	[in:play_music [sl:music_genre rappel]] [in:play_music [sl:music_artist_name rappel]] [in:create_reminder]
	ajoute l' enfant à l' appel	mBERT AMBER DAMP	[in:send_message [sl:recipient ajoute] [sl:content_exact à appel]] [in:update_call [sl:contact_added [in:get_contact [sl:contact_related l' enfant]]]] [in:update_call [sl:contact_added [in:get_contact [sl:type_relation enfant]]]]
	s' il te plait appelle adam .	mBERT AMBER DAMP	[in:question_music [sl:music_provider_name te adam]] [in:get_info_contact [sl:contact adam]] [in:create_call [sl:contact adam]]
	veuillez appeler peter	mBERT AMBER DAMP	[in:is_true_recipes [sl:recipes_dish veuillez peter]] [in:create_call [sl:contact veuillez peter]] [in:create_call [sl:contact peter]]
	veuillez appeler nick	mBERT AMBER DAMP	[in:question_news [sl:news_topic veuillez nick]] [in:get_contact [sl:contact veuillez nick]] [in:create_call [sl:contact nick]]
	efface toutes mes alarmes	mBERT AMBER DAMP	[in:update_alarm [sl:alarm_name efface mes]] [in:silence_alarm [sl:amount toutes]] [in:delete_alarm [sl:amount toutes]]
	peux - tu appeler amy	mBERT AMBER DAMP	[in:is_true_recipes [sl:recipes_dish peux tu] [sl:recipes_included_ingredient appeler amy]] [in:send_message [sl:recipient peux amy]] [in:create_call [sl:contact amy]]
	mets un réveil maintenant	mBERT AMBER DAMP	[in:get_timer [sl:contact mets]] [in:get_sunrise] [in:create_alarm]
	obtenez - moi des nouvelles	mBERT AMBER DAMP	[in:question_news [sl:news_topic obtenez nouvelles]] [in:get_stories_news [sl:news_type obtenez nouvelles]] [in:get_stories_news [sl:news_type nouvelles]]
	dis - moi quel temps il fait	mBERT AMBER DAMP	[in:get_info_recipes [sl:recipes_qualifier_nutrition dis fait]] [in:get_timer [sl:method_timer temps]] [in:get_weather]
	je dois appeler dave	mBERT AMBER DAMP	[in:question_news [sl:news_topic je dave]] [in:send_message [sl:recipient je dave]] [in:create_call [sl:contact dave]]
	merci d' appeler jessica	mBERT AMBER DAMP	[in:is_true_recipes [sl:recipes_dish merci jessica]] [in:update_call [sl:contact_added merci jessica]] [in:create_call [sl:contact jessica]]
	annule le rappel appeler maman	mBERT AMBER DAMP	[in:is_true_recipes [sl:recipes_attribute annule] [sl:recipes_included_ingredient rappel maman]] [in:update_call [sl:title_event annule maman]] [in:delete_reminder [sl:todo [in:create_call [sl:contact [in:get_contact [sl:type_relation maman]]]]]]
	ai - je reçu des appels de ma femme	mBERT AMBER DAMP	[in:question_news [sl:news_topic ai femme]] [in:get_call [sl:todo ai je] [sl:contact [in:get_contact [sl:contact_related ma] [sl:type_relation femme]]]] [in:get_call [sl:contact [in:get_contact [sl:contact_related ma] [sl:type_relation femme]]]]
	je voulais appeller edward weiss	mBERT AMBER DAMP	[in:question_news [sl:news_topic je weiss]] [in:get_info_contact [sl:contact je weiss]] [in:create_call [sl:contact edward weiss]]
	annule l' appel s' il te plait	mBERT AMBER DAMP	[in:question_news [sl:news_topic annule plait]] [in:question_news [sl:news_topic annule plait]] [in:end_call]
	quand maman m' a - t - elle appelé ?	mBERT AMBER DAMP	[in:question_news [sl:news_topic quand elle]] [in:get_call_time [sl:contact [in:get_contact [sl:type_relation maman]] [sl:contact m' elle]]] [in:get_call_time [sl:contact [in:get_contact [sl:type_relation maman]]]]

Table 6: Full Table of 20 Sampled French Results from Qualitative Analysis.

	Input	Model	Outputs
German	bbc - schlagzeilen	mBERT AMBER DAMP	[in:play_music [sl:music_artist_name bbc schlagzeilen]] [in:get_stories_news [sl:news_source bbc schlagzeilen]] [in:get_stories_news [sl:news_source bbc] [sl:news_type schlagzeilen]]
	erinnerung an urlaub	mBERT AMBER DAMP	[in:get_language [sl:contact urlaub]] [in:create_alarm [sl:alarm_name urlaub]] [in:create_reminder [sl:todo urlaub]]
	kannst du bitte meine mutter anrufen ?	mBERT AMBER DAMP	[in:get_recipes] [in:get_call [sl:category_event mutter]] [in:create_call [sl:contact [in:get_contact [sl:contact_related meine] [sl:type_relation mutter]]]]
	bitte schick die gruppe der frauen	mBERT AMBER DAMP	[in:question_news [sl:news_topic bitte frauen]] [in:get_info_recipes [sl:contact bitte schick] [sl:recipes_cuisine frauen]] [in:send_message [sl:group frauen]]
	rufe jeffrey whatsapp an	mBERT AMBER DAMP	[in:get_stories_news [sl:news_topic rufe jeffrey] [sl:name_app whatsapp]] [in:get_lyrics_music [sl:contact rufe jeffrey] [sl:name_app whatsapp]] [in:create_call [sl:contact jeffrey] [sl:name_app whatsapp]]
	wir rufen vincent roberts an	mBERT AMBER DAMP	[in:is_true_recipes [sl:recipes_included_ingredient wir roberts]] [in:get_info_contact [sl:contact vincent roberts]] [in:create_call [sl:contact vincent roberts]]
	spiel 98.9 radio auf iheartradio	mBERT AMBER DAMP	[in:play_music [sl:music_radio_id 98.9 auf] [sl:music_provider_name iheartradio]] [in:play_music [sl:music_radio_id spiel 98.9] [sl:music_type radio] [sl:music_provider_name iheartradio]] [in:play_music [sl:music_radio_id 98.9] [sl:music_type radio] [sl:music_provider_name iheartradio]]
	wen kenne ich in rice lake ?	mBERT AMBER DAMP	[in:get_education_time [sl:contact wen ich] [sl:location rice lake]] [in:get_location [sl:contact kenne ich] [sl:location rice lake]] [in:get_contact [sl:contact_related ich] [sl:location rice lake]]
	lancez l' appel à kelly	mBERT AMBER DAMP	[in:send_message [sl:recipient lancez kelly]] [in:create_call [sl:contact lancez kelly]] [in:create_call [sl:contact kelly]]
	ruf meine mutter an	mBERT AMBER DAMP	[in:send_message [sl:recipient ruf mutter]] [in:get_reminder [sl:alarm_name mutter]] [in:create_call [sl:contact [in:get_contact [sl:contact_related meine] [sl:type_relation mutter]]]]
	zeige politische nachrichten	mBERT AMBER DAMP	[in:send_message [sl:recipient zeige nachrichten]] [in:get_stories_news [sl:contact zeige] [sl:news_category politische] [sl:news_type nachrichten]] [in:get_stories_news [sl:news_category politische] [sl:news_type nachrichten]]
	rufe lucas an	mBERT AMBER DAMP	[in:create_reminder [sl:todo rufe lucas]] [in:play_media [sl:music_artist_name rufe an]] [in:create_call [sl:contact lucas]]
	wie macht man ropa vieja ?	mBERT AMBER DAMP	[in:get_contact [sl:contact macht ropa]] [in:get_info_contact [sl:contact ropa vieja]] [in:get_recipes [sl:recipes_dish ropa vieja]]
	rufe stattdessen nicole an	mBERT AMBER DAMP	[in:get_stories_news [sl:news_topic rufe nicole]] [in:play_media [sl:music_artist_name nicole an]] [in:create_call [sl:contact nicole]]
	ruf bitte henry an	mBERT AMBER DAMP	[in:create_timer [sl:contact ruf henry]] [in:play_media [sl:music_artist_name ruf an]] [in:create_call [sl:contact henry]]
	setze den timer jetzt fort	mBERT AMBER DAMP	[in:pause_timer [sl:method_timer timer]] [in:delete_timer [sl:method_timer timer]] [in:resume_timer [sl:method_timer timer]]
	bitte zeig mir alle alarme an	mBERT AMBER DAMP	[in:update_alarm [sl:alarm_name zeig alle]] [in:create_alarm [sl:amount alle]] [in:get_alarm [sl:amount alle]]
	beende den back - timer	mBERT AMBER DAMP	[in:update_timer [sl:method_timer timer]] [sl:timer_name back timer]] [in:pause_timer [sl:method_timer timer]]
	für wen arbeitet jerry ?	mBERT AMBER DAMP	[in:get_recipes [sl:recipes_attribute wen] [sl:recipes_dish arbeitet jerry]] [in:get_employer [sl:employer wen] [sl:contact jerry]] [in:get_employer [sl:contact jerry]]
	ist es fast fertig ?	mBERT AMBER DAMP	[in:get_stories_news [sl:news_source es fertig]] [in:get_weather [sl:weather_attribute fast fertig]] [in:get_timer]

Table 7: Full Table of 20 Sampled German Results from Qualitative Analysis.

	Input	Model	Outputs
Hindi	क्या कुछ हो रहा है	mBERT AMBER DAMP	[in:send_message [sl:recipient हो रहा]] [in:get_details_news] [in:get_event]
	फेसबुक में कौन काम करता है	mBERT AMBER DAMP	[in:question_news [sl:news_topic कौन करता]] [in:get_employer [sl:contact फेसबुक कौन]] [in:get_contact [sl:employer फेसबुक]]
	अंतिम कॉल किस समय किया गया था ?	mBERT AMBER DAMP	[in:question_news [sl:news_topic कॉल था]] [in:get_call_time [sl:contact अंतिम कॉल]] [in:get_call_time]
	मौसम सड़क को कैसे इफ़ेक्ट कर रहा है ?	mBERT AMBER DAMP	[in:get_stories_news [sl:news_topic मौसम रहा]] [in:question_news [sl:news_topic मौसम इफ़ेक्ट]] [in:get_weather]
	समर मोंटोया की दादी को कॉल करो	mBERT AMBER DAMP	[in:question_news [sl:news_topic समर करो]] [in:get_contact [sl:contact_related समर मोंटोया] [sl:type_relation दादी]]] [in:create_call [sl:contact [in:get_contact [sl:contact_related समर मोंटोया] [sl:type_relation दादी]]]]
	अलार्म्स बंद करें	mBERT AMBER DAMP	[in:create_reminder] [in:end_call [sl:contact अलार्म्स]] [in:silence_alarm]
	मेरे पापा को फ़ोन कॉल करो	mBERT AMBER DAMP	[in:is_true_recipes [sl:recipes_dish पापा करो]] [in:update_call [sl:contact [in:get_contact [sl:contact_related मेरे] [sl:type_relation पापा]]]] [in:create_call [sl:contact [in:get_contact [sl:contact_related मेरे] [sl:type_relation पापा]]]]
	रोपा विएजा को कैसे बनाया जाता है	mBERT AMBER DAMP	[in:question_news [sl:news_topic रोपा जाता]] [in:question_news [sl:location रोपा विएजा]] [in:get_recipes [sl:recipes_dish रोपा विएजा]]
	शाम को 6 बजे तापमान कैसा रहेगा ?	mBERT AMBER DAMP	[in:play_music [sl:music_album_title शाम तापमान]] [in:get_weather [sl:date_time शाम]] [in:get_weather [sl:date_time शाम बजे]]
	कृपया मेरे सभी अलार्म हटाएं ।	mBERT AMBER DAMP	[in:get_stories_news [sl:news_topic सभी]] [in:update_call [sl:contact_related मेरे] [sl:amount सभी]] [in:delete_alarm [sl:amount सभी]]
	क्या इस सोमवार कुछ मज़ेदार होने वाला है	mBERT AMBER DAMP	[in:send_message [sl:content_exact इस वाला]] [in:create_alarm [sl:date_time इस सोमवार]] [in:get_event [sl:date_time इस सोमवार]]
	अभी अलार्म शुरू करो	mBERT AMBER DAMP	[in:get_timer [sl:contact अभी करो]] [in:get_call [sl:contact अलार्म]] [in:create_alarm]
	मेरे लिए लमार को एक कॉल करो ।	mBERT AMBER DAMP	[in:question_news [sl:news_topic लिए करो]] [in:create_call [sl:contact मेरे लमार]] [in:create_call [sl:contact लमार]]
	मुझे ताज़ा ख़बरें बताओ	mBERT AMBER DAMP	[in:create_call [sl:contact बताओ]] [in:get_contact [sl:contact मुझे बताओ]] [in:get_stories_news [sl:date_time ताज़ा] [sl:news_type ख़बरें]]
	कृपया इसी समय ट्रेंट यू को कॉल करें	mBERT AMBER DAMP	[in:get_event [sl:category_event इसी समय] [sl:location को कॉल]] [in:question_news [sl:news_topic ट्रेंट करें]] [in:create_call [sl:contact ट्रेंट यू]]
	दोबारा बारिश कब हो सकती है	mBERT AMBER DAMP	[in:question_news [sl:news_topic दोबारा सकती]] [in:send_message [sl:recipes_dish दोबारा बारिश]] [in:get_weather [sl:weather_attribute बारिश]]
	जोसेफ नंबर दो को कॉल करें	mBERT AMBER DAMP	[in:send_message [sl:recipient दो कॉल]] [in:get_availability [sl:contact जोसेफ]] [in:create_call [sl:contact जोसेफ] [sl:ordinal दो]]
	मेरा टाइमर वापस शुरू करें	mBERT AMBER DAMP	[in:send_message [sl:recipient टाइमर वापस]] [in:update_reminder_date_time [sl:todo मेरा टाइमर]] [in:resume_timer [sl:method_timer टाइमर]]
	सैल क्रोकेट को कॉल करने की कोशिश करो	mBERT AMBER DAMP	[in:send_message [sl:content_exact को करो]] [in:send_message [sl:recipient सैल क्रोकेट]] [in:create_call [sl:contact सैल क्रोकेट]]
	वर्तमान गीत दोहराएं	mBERT AMBER DAMP	[in:create_call [sl:contact गीत]] [in:loop_music [sl:music_type गीत]] [in:replay_music [sl:music_type गीत]]

Table 8: Full Table of 20 Sampled Hindi Results from Qualitative Analysis.

	Input	Model	Outputs
Thai	โทร หา สตีเวน	mBERT AMBER DAMP	[in:play_music [sl:music_artist_name หา]] [in:get_contact [sl:contact โทร สตีเวน]] [in:create_call [sl:contact สตีเวน]]
	ช่วย เริ่ม ต้น การ โทร กับ จู เนียร์	mBERT AMBER DAMP	[in:get_event [sl:location การ]] [in:update_call [sl:contact_added ช่วย โทร] [sl:music_artist_name จู เนียร์]] [in:create_call [sl:contact จู เนียร์]]
	วาง สาย	mBERT AMBER DAMP	[in:get_info_recipes [sl:content_exact วาง สาย]] [in:update_call [sl:contact วาง สาย]] [in:end_call]
	เชิญ เบธ เข้า ร่วม การ โทร นี้	mBERT AMBER DAMP	[in:get_event [sl:location การ]] [in:question_news [sl:news_topic เชิญ นี้]] [in:update_call [sl:contact_added เบธ]]
	ตั้ง เตือน ความ จำ เพื่อ ล้าง สี ผม	mBERT AMBER DAMP	[in:send_message [sl:content_exact ความ จำ]] [in:question_news [sl:news_topic ตั้ง ผม]] [in:create_reminder [sl:todo ล้าง ผม]]
	ตั้ง นาฬิกา ปลุก ใน 20 นาที	mBERT AMBER DAMP	[in:create_alarm [sl:date_time 20]] [in:create_alarm [sl:alarm_name นาฬิกา ปลุก] [sl:date_time ใน นาที]] [in:create_alarm [sl:date_time ใน นาที]]
	ใคร ทำงาน ที่ at & t	mBERT AMBER DAMP	[in:send_message [sl:recipient ทำงาน] [sl:todo at t]] [in:get_contact [sl:contact ใคร ทำงาน] [sl:employer at t]] [in:get_contact [sl:employer at t]]
	โทร หา เดย์ และ โฮสท์	mBERT AMBER DAMP	[in:play_music [sl:music_artist_name หา]] [in:play_media [sl:music_artist_name โทร เดย์] [sl:recipes_attribute โฮสท์]] [in:create_call [sl:contact เดย์] [sl:contact โฮสท์]]
	วาง สาย ฌอน	mBERT AMBER DAMP	[in:get_stories_news [sl:location วาง สาย]] [in:play_media [sl:music_artist_name วาง ฌอน]] [in:end_call [sl:contact ฌอน]]
	มี ใคร ว่าง ไหม	mBERT AMBER DAMP	[in:send_message [sl:recipient ไหม]] [in:get_info_contact [sl:contact ใคร ว่าง]] [in:get_availability]
	อากาศ จะ เป็น อย่างไร พรุ่ง นี้ เช้า ?	mBERT AMBER DAMP	[in:create_alarm] [in:question_news [sl:news_topic อากาศ เช้า]] [in:get_weather [sl:date_time พรุ่ง เช้า]]
	ฉัน ต้อง โทร หา เดฟ	mBERT AMBER DAMP	[in:send_message [sl:recipient หา]] [in:send_message [sl:recipient ฉัน เดฟ]] [in:create_call [sl:contact เดฟ]]
	จะ มี พายุ ใน วัน เสาร์ หรือ ไม่ ?	mBERT AMBER DAMP	[in:get_event] [in:get_weather [sl:alarm_name มี พายุ] [sl:date_time ใน เสาร์] [sl:date_time ไม่]] [in:get_weather [sl:weather_attribute พายุ] [sl:date_time ใน เสาร์]]
	อุณหภูมิ ข้าง นอก เป็น อย่างไร	mBERT AMBER DAMP	[in:send_message [sl:recipient นอก]] [in:question_news [sl:news_topic อุณหภูมิ นอก]] [in:get_weather]
	ช่วย เริ่ม การ โทร หา พ่อ ให้ หน่อย	mBERT AMBER DAMP	[in:send_message [sl:recipient การ] [sl:content_exact หา]] [in:send_message [sl:todo ช่วย หน่อย]] [in:create_call [sl:contact [in:get_contact [sl:type_relation พ่อ]]]]
	สภาพ ภูมิ อากาศ ใน ฮาวาย เป็น อย่างไร	mBERT AMBER DAMP	[in:get_event] [in:question_news [sl:news_topic สภาพ ฮาวาย]] [in:get_weather [sl:location ฮาวาย]]
	ใน สาม วัน สภาพ อากาศ จะ เป็น อย่างไร ?	mBERT AMBER DAMP	[in:send_message [sl:recipient สาม]] [in:question_news [sl:date_time ใน อย่างไร]] [in:get_weather [sl:date_time ใน วัน]]
	เตือน ความ จำ ใหม่	mBERT AMBER DAMP	[in:send_message [sl:recipient ความ จำ]] [in:create_reminder [sl:person_reminded ความ ใหม่]] [in:create_reminder]
	โทร ไป ที่ 5405551560	mBERT AMBER DAMP	[in:get_event [sl:location ไป 5405551560]] [in:create_call [sl:phone_number โทร 5405551560]] [in:create_call [sl:phone_number 5405551560]]
	มี การ จัด คอนเสิร์ต อะไร บ้าง	mBERT AMBER DAMP	[in:play_music [sl:music_artist_name การ]] [in:question_news [sl:news_topic การ บ้าง]] [in:get_event [sl:category_event คอนเสิร์ต]]

Table 9: Full Table of 20 Sampled Thai Results from Qualitative Analysis.