Optimal Transport Posterior Alignment for Cross-lingual Semantic Parsing

Tom Sherborne and **Tom Hosking** and **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
{tom.sherborne, tom.hosking}@ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

Cross-lingual semantic parsing transfers parsing capability from a high-resource language (e.g., English) to low-resource languages with scarce training data. Previous work has primarily considered silver-standard data augmentation or zero-shot methods; exploiting few-shot gold data is comparatively unexplored. We propose a new approach to cross-lingual semantic parsing by explicitly minimizing cross-lingual divergence between probabilistic latent variables using Optimal Transport. We demonstrate how this direct guidance improves parsing from natural languages using fewer examples and less training. We evaluate our method on two datasets, MTOP and MultiATIS++SQL, establishing state-of-the-art results under a few-shot cross-lingual regime. Ablation studies further reveal that our method improves performance even without parallel input translations. In addition, we show that our model better captures cross-lingual structure in the latent space to improve semantic representation similarity.[1]

1 Introduction

Semantic parsing maps natural language utterances to logical form (LF) representations of meaning. As an interface between human- and computer-readable languages, semantic parsers are a critical component in various natural language understanding (NLU) pipelines, including assistant technologies (Kollar et al., 2018), knowledge base question answering (Berant et al., 2013; Liang, 2016), and code generation (Wang et al., 2023).

Recent advances in semantic parsing have led to improved reasoning over challenging questions (Li et al., 2023) and accurate generation of complex queries (Scholak et al., 2021), however, most prior work has focused on English (Kamath and Das, 2019; Qin et al., 2022a). Expanding, or *localizing*, an English-trained model to additional languages is challenging for several reasons. There is typically little labeled data in the target languages due to high annotation costs. Cross-lingual parsers must also be sensitive to how different languages refer to entities or model abstract and mathematical relationships (Reddy et al., 2017; Hershcovich et al., 2019). Transfer between dissimilar languages can also degrade in multilingual models with insufficient capacity (Pfeiffer et al., 2022).

Previous strategies for resource-efficient localization include generating ‘‘silver-standard’’ training data through machine-translation (Nicosia et al., 2021) or prompting large language models (Rosenbaum et al., 2022). Alternatively, zero-shot models use ‘‘gold-standard’’ external corpora for auxiliary tasks (van der Goot et al., 2021) and few-shot models maximize sample-efficiency using meta-learning (Sherborne and Lapata, 2023). We argue that previous work encourages cross-lingual transfer through *implicit* alignment only via minimizing silver-standard data perplexity, multi-task ensembling, or constraining gradients.

We instead propose to localize an encoder-decoder semantic parser by *explicitly* inducing cross-lingual alignment between representations. We present MINOTAUR (**Min**imizing **O**ptimal **T**ransport distance for **A**lignment **U**nder **R**epresentations)—a method for cross-lingual semantic parsing which explicitly minimizes distances between probabilistic latent variables to reduce representation divergence across languages (Figure 1). MINOTAUR leverages Optimal Transport theory (Villani, 2008) to measure and minimize this divergence between English and target languages during episodic few-shot learning. Our hypothesis is that

[1] Our code and data are publicly available at github.com/tomsherborne/minotaur.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 1432–1450, 2023. https://doi.org/10.1162/tacl_a_00611
Action Editor: Xavier Carreras. Submission batch: 4/2023; Revision batch: 5/2023; Published 11/2023.

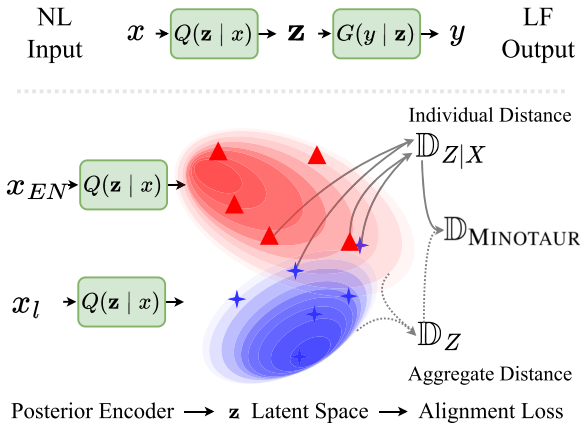

Figure 1: **Upper**: We align representations explicitly in the latent representation space, $\mathbf{z}$, between encoder Q and decoder G. **Lower**: MINOTAUR induces cross-lingual similarity by minimizing divergence between latent distributions at two levels—between individual and aggregate posteriors.

explicit alignment between latent variables can improve knowledge transfer between languages without requiring additional annotations or lexical alignment. We evaluate this hypothesis in a *few-shot* cross-lingual regime and study how many examples in languages beyond English are needed for ''good'' performance.

Our technique allows us to precisely measure, and minimize, the cross-lingual transfer gap between languages. This yields both sample-efficient training and establishes leading performance for few-shot cross-lingual transfer on two datasets. We focus our evaluation on semantic parsing but MINOTAUR can be applied directly to a wide range of other tasks. Our contributions are as follows:

- We propose a method for learning a semantic parser using *explicit* cross-lingual alignment between probabilistic latent variables. MINOTAUR jointly minimizes marginal and conditional posterior divergence for *fast* and *sample-efficient* cross-lingual transfer.
- We propose an episodic training scheme for cross-lingual posterior alignment during training which requires minimal modifications to typical learning.
- Experiments on task-oriented semantic parsing (MTOP; Li et al., 2021) and executable semantic parsing (MultiATIS++SQL; Sherborne and Lapata, 2022) demonstrate that MINOTAUR outperforms prior methods with fewer data resources and faster convergence.

2 Related Work

Cross-lingual Semantic Parsing Growing interest in cross-lingual NLU has motivated the expansion of benchmarks to study model adaptation across many languages (Hu et al., 2020; Liang et al., 2020). Within executable semantic parsing, ATIS (Hemphill et al., 1990) has been translated into multiple languages such as Chinese and Indonesian (Susanto and Lu, 2017a), and GeoQuery (Zelle and Mooney, 1996) has been translated into German, Greek, and Thai (Jones et al., 2012). Adjacent research in Task-Oriented Spoken Language Understanding (SLU) has given rise to datasets such as MTOP in five languages (Li et al., 2021), and MultiATIS++ in seven languages (Xu et al., 2020). SLU aims to parse inputs into functional representations of dialog acts (which are often embedded in an assistant NLU pipeline) instead of executable machine-readable language.

In all cases, cross-lingual semantic parsing demands fine-grained semantic understanding for successful transfer across languages. Multilingual pre-training (Pires et al., 2019) has the potential to unlock certain understanding capabilities but is often insufficient. Previous methods resort to expensive dataset translation (Jie and Lu, 2014; Susanto and Lu, 2017b) or attempt to mitigate data paucity by creating ''silver'' standard data through machine translation (Sherborne et al., 2020; Nicosia et al., 2021; Xia and Monti, 2021; Guo et al., 2021) or prompting (Rosenbaum et al., 2022; Shi et al., 2022). However, methods that rely on synthetic data creation are yet to produce cross-lingual parsing equitable to using gold-standard professional translation.

Zero-shot methods bypass the need for in-domain data augmentation using multi-task objectives which incorporate gold-standard data for external tasks such as language modeling or dependency parsing (van der Goot et al., 2021; Sherborne and Lapata, 2022; Gritta et al., 2022). Few-shot approaches which leverage a small number of annotations have shown promise in various tasks (Zhao et al., 2021, *inter alia*) including semantic parsing. Sherborne and Lapata (2023) propose a first-order meta-learning algorithm to train

a semantic parser capable of sample-efficient cross-lingual transfer.

Our work is most similar to recent studies on cross-lingual alignment for classification tasks (Wu and Dredze, 2020) and spoken-language understanding using token- and slot-level annotations between parallel inputs (Qin et al., 2022b; Liang et al., 2022). While similar in motivation, we contrast in our exploration of latent variables with parametric alignment for a closed-form solution to cross-lingual transfer. Additionally, our method does not require fine-grained word and phrase alignment annotations, instead inducing alignment in the continuous latent space.

Alignment and Optimal Transport Optimal Transport (OT; Villani, 2008) minimizes the cost of mapping from one distribution (e.g., utterances) to another (e.g., logical forms) through some joint distribution with conditional independence (Monge, 1781), i.e., a latent variable conditional on samples from one input domain. OT in NLP has mainly used Sinkhorn distances to measure the divergence between non-parametric discrete distributions as an online minimization subproblem (Cuturi, 2013).

Cross-lingual approaches to OT have been proposed for embedding alignment (Alvarez-Melis and Jaakkola, 2018; Alqahtani et al., 2021), bilingual lexicon induction (Marchisio et al., 2022), and summarization (Nguyen and Luu, 2022). Our method is similar to recent proposals for cross-lingual retrieval using variational or OT-oriented representation alignment (Huang et al., 2023; Wieting et al., 2023). Wang and Wang (2019) consider a ''continuous'' perspective on OT using the Wasserstein Auto-Encoder (Tolstikhin et al., 2018, WAE) as a language model which respects geometric input characteristics within the latent space.

Our parametric formulation allows this continuous approach to OT, similar to the WAE model. While monolingual prior work in semantic parsing has identified that latent structure can benefit the semantic parsing task (Kočiský et al., 2016; Yin et al., 2018), it does not consider whether it can inform transfer between languages. To the best of our knowledge, we are the first to consider the continuous form of OT for cross-lingual transfer in a sequence-to-sequence task. We formulate the parsing task as a transportation problem in Section 3 and describe how this framework gives rise to explicit cross-lingual alignment in Section 4.

3 Background

3.1 Cross-lingual Semantic Parsing

Given a natural language utterance x, represented as a sequence of tokens $(x_1, \ldots, x_T)$, a semantic parser generates a faithful logical-form meaning representation y.[2] A typical neural network parser trains on input-output pairs $\{x_i,\ y_i\}_{i=0}^{N}$, using the cross-entropy between predicted $\hat{y}$, and gold-standard logical form y, as supervision (Cheng et al., 2019).

Following the standard VAE framework (Kingma and Welling, 2014; Rezende et al., 2014), an encoder Q_ϕ represents inputs from $\mathcal{X}$ as a continuous latent variable Z, $Q_\phi : \mathcal{X} \rightarrow \mathcal{Z}$. A decoder G_θ predicts outputs conditioned on samples from the latent space, $G_\theta : \mathcal{Z} \rightarrow \mathcal{Y}$. The encoder therefore acts as approximate posterior $Q_\phi(Z|X)$. Q_ϕ is a multi-lingual pre-trained encoder shared across all languages.

For cross-lingual transfer, the parser must also generalize to languages from which it has seen few (or zero) training examples.[3] Our goal is for the prediction for input $x_l \in X_l$ in language l to match the prediction for equivalent input from a high-resource language (typically English), i.e., $x_l \rightarrow y,\ x_{\text{EN}} \rightarrow y$ subject to the constraint of fewer training examples in l ($|N_l| \ll |N_{\text{EN}}|$). As shown in Figure 1, we propose measuring the divergence between approximate posteriors (i.e., $Q\left(Z|X_{\text{EN}}\right)$ and $Q\left(Z|X_l\right)$) as the distance between individual samples and an approximation of the ''mean'' encoding of each language. This goal of aligning distributions naturally fits an Optimal Transport perspective.

3.2 Kantorovich Transportation Problem

Tolstikhin et al. (2018) propose the Wasserstein Auto-Encoder (WAE) as an alternative variational model. The WAE minimizes the *transportation cost* under the Kantorovich form of the Optimal Transport problem (Kantorovich, 1958). Given two distributions P_X, P_Y, the objective is to find a *transportation plan* $\Gamma\left(X,\ Y\right)$, within the set of

[2]Notation key: Capitals X, are random variables; Curly $\mathcal{X}$, are functional domains; lowercase x are observations and $P_{\{\}}$ are probability distributions.

[3]Resource parity between languages is *multilingual* semantic parsing which we view as an upper-bound.

all joint distributions, $\mathcal{P}(X \sim P_X,\ Y \sim P_Y)$, to map probability mass from P_X to P_Y with minimal cost. T_c expresses the problem of finding a plan which minimizes a transportation cost function $c(X,\ Y) :\ \mathcal{X} \times \mathcal{Y} \to \mathcal{R}_+$:

$$T_c(P_X, P_Y) := \inf_{\Gamma \in (X \sim P_X,\ Y \sim P_Y)} \mathbb{E}_{(X,Y)\sim\Gamma}\left[c(X,\ Y)\right] \tag{1}$$

The WAE is proposed as an auto-encoder (i.e., P_Y approximates P_X), however, in our setting P_X is the natural language input distribution and P_Y is the logical form output distribution and they are both realizations of the same semantics.

Using conditional independence, $y \perp\!\!\!\perp x \mid \mathbf{z}$, we can transform the plan, $\Gamma(X,\ Y) \to \Gamma(Y|X)\, P_X$ and consider a non-deterministic mapping from X to Y under observed P_X. Tolstikhin et al. (2018, Theorem 1) identify how to *factor* this mapping through latent variable Z, leading to:

$$T_c(P_X, P_Y) = \inf_{Q_\phi(Z|X)\in\mathcal{Q}} \mathbb{E}_{P_X}\mathbb{E}_{Q_\phi(Z|X)}\left[c(Y, G_\theta(Z))\right] + \alpha\mathbb{D}\left(Q(Z), P(Z)\right) \tag{2}$$

Equation (2) expresses a minimizable objective: identify the probabilistic encoder $Q_\phi(Z|X)$ and decoder $G_\theta(Z)$ which minimizes a cost, subject to regularization on the divergence $\mathbb{D}$ between the marginal posterior $Q(Z)$ and prior $P(Z)$.

The additional regularization is how the WAE improves on the evidence lower bound in the variational auto-encoder, where the equivalent alignment on the individual posterior $Q_\phi(Z|X)$ drives latent representations to zero. Regularization on the marginal posterior $Q(Z) = \mathbb{E}_{X\sim P_X}\left[Q_\phi(Z|X)\right]$ instead allows individual posteriors for different samples to remain distinct and non-zero. This limits posterior collapse, guiding Z to remain informative for decoding.

We use Maximum Mean Discrepancy (Gretton et al., 2012, MMD) for an unbiased estimate of $\mathbb{D}(Q(Z), P(Z))$ as a robust measure of the distance between high dimensional Gaussian distributions. Equation (3) defines MMD using some kernel $k :\ \mathcal{Z} \times \mathcal{Z} \to \mathcal{R}$, defined over a reproducible kernel Hilbert space, $\mathcal{H}_k$:

$$\text{MMD}_k(P,\ Q) = \left\| \int_{\mathcal{Z}} k(z, \cdot)\, dP - \int_{\mathcal{Z}} k(z, \cdot)\, dQ \right\|_{\mathcal{H}_k} \tag{3}$$

Informally, MMD minimizes the distance between the ''feature means'' of variables P and Q estimated over a batch sample. Equation (4) defines MMD estimation over observed $\mathbf{p}$ and $\mathbf{q}$ using the heavy-tailed *inverse multiquadratic* (IMQ) kernel k:

$$\text{MMD}_k(\mathbf{p}, \mathbf{q}) = \frac{1}{n_p(n_p-1)} \sum_{z'\neq z} k(p_z,\ p_{z'}) + \tag{4}$$

$$\frac{1}{n_q(n_q-1)} \sum_{z'\neq z} k(q_z,\ q_{z'}) - \frac{2}{n_p n_q} \sum_{z,\ z'} k(p_z,\ q_{z'})$$

We define the IMQ kernel in Equation (5) below; $C = 2|\mathbf{z}|\sigma^2$ and $\mathcal{S} = [0.1, 0.2, 0.5, 1, 2, 5, 10]$.

$$k(p, q) = \sum_{s\in\mathcal{S}} \frac{s \cdot C}{s \cdot C + \|p - q\|_2^2} \tag{5}$$

This framework defines a WAE objective using a cost function, c to map from P_X to P_Y through latent variable Z. We now describe how MINOTAUR integrates explicit posterior alignment during this learning process.

4 MINOTAUR: Posterior Alignment for Cross-lingual Transfer

Variational Encoder-Decoder Our model comprises of encoder (and approximate posterior) Q_ϕ, and generator decoder G_θ. The encoder Q_ϕ produces a distribution over latent encodings $\mathbf{z} = \{z_1, \ldots,\ z_T\}$, parameterized as a sequence of T mean states $\boldsymbol{\mu}_{\{1,\ldots,T\}} \in \mathbb{R}^{T\times d}$, and a single variance $\sigma^2 \in \mathbb{R}^d$ for all T states,

$$\mathbf{z} = Q_\phi(x) \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2) \tag{6}$$

The latent encodings $\mathbf{z}$ are sampled using the Gaussian reparameterization trick (Kingma and Welling, 2014),

$$\mathbf{z} = \boldsymbol{\mu} + \sigma^2 \circ \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{7}$$

Finally, an output sequence $\hat{y}$ is generated from $\mathbf{z}$ through autoregressive generation,

$$\hat{y} = G_\theta(\mathbf{z}) \tag{8}$$

For an input sequence of T tokens, we use a sequence of T latent variables for $\mathbf{z}$ over pooling into a single representation. This allows for

more 'bandwidth' in the latent state to minimize the risk of the decoder ignoring $\mathbf{z}$, i.e., *posterior collapse*. We find this design choice to be necessary as lossy pooling leads to weak overall performance. We also use a single variance estimate for sequence $\mathbf{z}$—this minimizes variance noise across $\mathbf{z}$ and simplifies computation in posterior alignment. We follow the convention of an isotropic unit Gaussian prior, $P(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Cross-lingual Alignment Typical WAE modeling builds meaningful latent structure by aligning the estimated posterior to the prior only. MINOTAUR extends this through additionally aligning posteriors *between* languages. Consider learning the optimal mapping from English utterances X_{EN} to logical forms Y within Equation (1) via latent variable Z, from monolingual data (X_{EN}, Y). The optimization in Equation (2) converges on an optimal transportation plan Γ^*_{EN} as the minimum cost.[4]

For transfer from English to language l, previous work either requires token alignment between X_{EN} and X_l or exploits the shared Y between X_{EN} and X_l (Qin et al., 2022b, *inter alia*). We instead induce alignment by explicitly matching Z between languages. Since Y is dependent only on Z, the latent variable offers a continuous representation space for alignment with the minimal and intuitive condition that equivalent z yields equivalent y. Therefore, our proposal is a straightforward extension of learning Γ^*_{EN}; we propose to bootstrap the transportation plan for target language l (i.e., $\Gamma^*_l(X_l, Y)$) by aligning on Z in a few-shot learning scenario. MINOTAUR *explicitly* aligns Z_l (from a target language l) towards Z (from EN) by matching $Q(Z_l|X_l)$ to $Q(Z|X_{\text{EN}})$ for the goal $\Gamma^*_l = \Gamma^*_{\text{EN}}$, thereby transferring the learned capabilities from high-resource languages with only a few training examples.

Given parallel inputs x_{EN} and x_l in English and language l, with equivalent LF ($y_{\text{EN}} = y_l$), their latent encodings are given by:

$$\mathbf{z}_{\text{EN}} = Q_\phi(x_{\text{EN}}), \hat{y}_{\text{EN}} = G(\mathbf{z}_{\text{EN}}) \quad (9)$$

$$\mathbf{z}_l = Q_\phi(x_l), \hat{y}_l = G(\mathbf{z}_l) \quad (10)$$

Unlike vanilla VAEs, where $\mathbf{z}$ is a single vector, the posterior samples ($\mathbf{z}_{\text{EN}}, \mathbf{z}_l \in \mathbb{R}^{T \times d}$) are complex structures. We therefore follow Mathieu et al. (2019) in using a decomposed alignment signal minimizing both *aggregate* posterior alignment (higher-level) and *individual* posterior alignment (lower-level) with scaling factors (α_P, β_P) respectively. This leads to the MINOTAUR alignment outlined in Figure 1 and expressed below,

$$\begin{aligned}\mathbb{D}_{\text{MINOTAUR}}(\mathbf{z}_{\text{EN}}, \mathbf{z}_l) = \\ \alpha_P \mathbb{D}_Z(Q_\phi(\mathbf{z}_{\text{EN}}), Q_\phi(\mathbf{z}_l)) \\ + \beta_P \mathbb{D}_{Z|X}(Q_\phi(\mathbf{z}_l|x_l) \| Q_\phi(\mathbf{z}_{\text{EN}}|x_{\text{EN}}))\end{aligned} \quad (11)$$

where $\mathbb{D}_{Z|X}$ is a divergence penalty between *individual* representations to match local structure, while $\mathbb{D}_Z$ is a divergence penalty between representation *aggregates* to match more global structure. The intuition is that individual matching promotes contextual encoding similarity and aggregate matching promotes similarity at the language level.

Similar to the prior alignment, we use the MMD distance to align aggregate posteriors as Equation (3) (i.e., marginal posteriors over Z between languages). For individual alignment, we consider two numerically stable *exact* solutions to measure individual divergence which are well suited to matching high-dimensional Gaussians (Takatsu, 2011). Modeling $Q_\phi(Z|X)$ as a parametric statistic yields the benefit of closed-form computation during learning. We primarily use the L^2 Wasserstein distance, W_2, as the Optimal Transport-derived minimum transportation cost between Gaussians $(\mathbf{p}, \mathbf{q})$ across domains. Within Equation (12) the mean is μ, covariance is $\Sigma = \text{Diag}\{\sigma_i^2, \ldots, \sigma_n^2\}$, and encodings have dimensionality d. Tr$\{\}$ is the matrix trace function.

$$\begin{aligned}W_2(\mathbf{p}, \mathbf{q}) = \|\mu_{\mathbf{p}} - \mu_{\mathbf{q}}\|_2^2 + \\ \text{Tr}\{\Sigma_{\mathbf{p}} + \Sigma_{\mathbf{q}} - 2\left(\Sigma_{\mathbf{p}}^{\frac{1}{2}} \Sigma_{\mathbf{q}} \Sigma_{\mathbf{p}}^{\frac{1}{2}}\right)^{\frac{1}{2}}\}\end{aligned} \quad (12)$$

We also consider the Kullback-Leibler Divergence (KL) between two Gaussian distributions as Equation (13). Minimizing KL is equivalent to maximizing the mutual information between distributions as an information-theoretic goal of semantically aligning $\mathbf{z}$. Section 6 demonstrates that W_2 is superior to KL in all cases.

$$\begin{aligned}\text{KL}(\mathbf{p}\|\mathbf{q}) = \frac{1}{2}\left(\log\left(\frac{|\Sigma_{\mathbf{q}}|}{|\Sigma_{\mathbf{p}}|}\right) - d_{p,q} + \right. \\ \left. \text{Tr}\{\Sigma_{\mathbf{q}}^{-1}\Sigma_{\mathbf{p}}\} + (\mu_{\mathbf{q}} - \mu_{\mathbf{p}})^T \Sigma_{\mathbf{q}} (\mu_{\mathbf{q}} - \mu_{\mathbf{p}})\right)\end{aligned} \quad (13)$$

[4] Γ_* is implicit within the model parameters.

Tr{} is the matrix trace function. Minimizing KL is equivalent to maximizing the mutual information between distributions as an information-theoretic goal of semantically aligning $\mathbf{z}$. Section 6 demonstrates that W_2 is superior to KL in all cases.

We express $\mathbb{D}_{Z|X}$ (see Equation (11)) between singular $\mathbf{p}$ and $\mathbf{q}$ representations for individual tokens for clarity, however, we actually minimize the *mean* of $\mathbb{D}_{Z|X}$ between each $\mathbf{z}_1$ and $\mathbf{z}_2$ tokens across both sequences, i.e., $\frac{1}{|\mathbf{z}_1||\mathbf{z}_2|}\sum_{i,j}\mathbb{D}_{Z|X}(\mathbf{z}_{1i}\|\mathbf{z}_{2j})$. We observe that minimizing this mean divergence between all $(\mathbf{z}_{1i}, \mathbf{z}_{2j})$ pairs is most empirically effective.

Finally, Equation (14) expresses the transportation cost, T_c, for a single $(x,\ y)$ pair during training: the cross-entropy between predicted and gold y and WAE marginal prior regularization.

$$\mathcal{L}(x,y) = \mathbb{E}_{Q(\mathbf{z}|x)}\left[-\sum_i y_i \left(\log G_\theta\left(\mathbf{z}\right)\right)_i\right] + \alpha\mathbb{D}\left(Q_\phi(\mathbf{z}), P(\mathbf{z})\right) \quad (14)$$

We episodically augment Equation (14) as Equation (15) using the MINOTAUR loss every k steps for few-shot induction of cross-lingual alignment. Sampling (x, y) is detailed in Section 5.

$$\mathcal{L}_\Sigma = \mathcal{L}(x_{\text{EN}}, y_{\text{EN}}) + \mathcal{L}(x_l, y_l) + \mathbb{D}_{\text{MINOTAUR}}\left(\mathbf{z}_{\text{EN}}, \mathbf{z}_l\right) \quad (15)$$

Another perspective on our approach is that we are aligning pushforward distributions, $Q(X) : \mathcal{X} \rightarrow \mathcal{Z}$. Cross-lingual alignment at the input token level (in $\mathcal{X}$) requires fine-grained annotations and is an outstanding research problem (see Section 2). Our method of aligning pushforwards in $\mathcal{Z}$ is smoothly continuous, does not require word alignment, and does not always require input utterances to be parallel translations. While we evaluate MINOTAUR principally on semantic parsing, our framework can extend to any sequence-to-sequence or representation learning task which may benefit from explicit alignment between languages or domains.

5 Experimental Setting

MTOP (Li et al., 2021) This contains dialog utterances of ‘‘assistant’’ queries and their corresponding tree-structured slot and intent LFs. MTOP is split into 15,667 training, 2,235 validation, and 4,386 test examples in English (EN).

x_{EN}	`word1` Who `word2` attended `word3` Yale?
x_{DE}	`word1` Wer `word2` besuchte `word3` Yale?
y	`[IN:GET_CONTACT [SL:SCHOOL word3 ]]`
x_{EN}	What does ORD mean?
x_{FR}	Que signifie ORD?
y	`SELECT DISTINCT airport.airport_name FROM airport WHERE airport.code=ORD;`

Figure 2: Input, x, and output, y, examples in English (EN), German (DE), and French (FR) for MTOP (Li et al., 2021, upper green) and MultiATIS++SQL (Sherborne and Lapata, 2022, lower red), respectively.

A variable subsample of each split is translated into French (FR), Spanish (ES), German (DE), and Hindi (HI). We refer to Li et al. (2021, Table 1) for complete dataset details. As shown in Figure 2, we follow Rosenbaum et al. (2022, Appendix B.2) using ‘‘space-joined’’ tokens and ‘‘sentinel words’’ (i.e., a `wordi` token is prepended to each input token and replaces this token in the LF) to produce a closed decoder vocabulary (Raman et al., 2022). This allows the output LF to reference input tokens by label without a copy mechanism. We evaluate LF accuracy using the *Space and Case Invariant Exact-Match* metric (SCIEM; Rosenbaum et al., 2022).

We sample a small number of training instances for low-resource languages, following the *Samples-per-Intent-and-Slot* (SPIS) strategy from Chen et al. (2020) which we adapt to our cross-lingual scenario. SPIS randomly selects examples and keeps those that mention any slot and intent value (e.g., ‘‘`IN:`’’ and ‘‘`SL:`’’ from Figure 2) with fewer than some rate in the existing subset. Sampling stops when all slots and intents have a minimum frequency of the sampling rate (or the maximum if fewer than the sampling rate). SPIS sampling ensures a minimum coverage of all slot and intent types during cross-lingual transfer. This normalizes unbalanced low-resource data as the model has seen approximately similar examples across all semantic categories. Practically, an SPIS rate of 1, 5, and 10 equates to 284 (1.8%), 1,125 (7.2%), and 1,867 (11.9%) examples (% training data).

MultiATIS++SQL (Sherborne and Lapata, 2022) Experiments on ATIS (Hemphill et al., 1990) study cross-lingual transfer using an executable LF to retrieve database information. We

use the MultiATIS++SQL version (see Table 2), pairing executable SQL with parallel inputs in English (EN), French (FR), Portuguese (PT), Spanish (ES), German (DE), and Chinese (ZH). We measure *denotation accuracy*—the proportion of executed predictions retrieving equivalent database results as executing the gold LF. Data is split into 4,473 training, 493 validation, and 448 test examples with complete translation for all splits. We follow Sherborne and Lapata (2023) in using random sampling. Rates of 1%, 5%, and 10% correspond to 45, 224, and 447 examples, respectively. For both datasets, the model only observes remaining data in English, e.g., sampling at 5% uses 224 multilingual examples and 4,249 English-only examples for training.

Modeling We follow prior work in using a Transformer encoder-decoder: We use the frozen pre-trained 12-layer encoder from mBART50 (Tang et al., 2021) and append an identical learnable layer. The decoder is a six-layer Transformer stack (Vaswani et al., 2017) matching the encoder dimensionality ($d = 1,024$). Decoder layers are trained from scratch following prior work and early experiments verified that pre-training the decoder did not assist in cross-lingual transfer, offering minimal improvement on English. The variance predictor (σ^2 for predicting $\mathbf{z}$ in Equation (6)) is a multi-head pooler from Liu and Lapata (2019) adapting multi-head attention to produce singular output from sequential inputs. The final model has $\sim$116 million trainable parameters and $\sim$340 million frozen parameters.

Optimization We train for a maximum of ten epochs with early stopping using validation loss. Optimization uses Adam (Kingma and Ba, 2015) with a batch size of 256 and learning rate of 1×10^{-4}. We empirically tune hyperparameters (β_P, α_P) to $(0.5, 0.01)$, respectively. During learning, a typical step (without Minotaur alignment) samples a batch of (x_L, y) pairs in languages $L \in \{\text{EN}, l_1, l_2 \ldots\}$ from a sampled dataset described above. Each Minotaur step instead uses a sampled batch of parallel data $(x_{\text{EN}},\ x_l,\ y_{\text{EN}},\ y_l)$ to induce explicit cross-lingual alignment from the same data pool. The episodic learning loop size is tuned to $k = 20$; we find that if k is infrequent then posterior alignment is weaker and if k is too frequent then overall parsing degrades as the posterior alignment dominates learning. Tokenization uses SentencePiece (Kudo and Richardson, 2018) and beam search prediction uses five hypotheses. All experiments are implemented in PyTorch (Paszke et al., 2019) and AllenNLP (Gardner et al., 2018). Training takes one hour using 1× A100 80GB GPU for either dataset.

Comparison Systems As an upper-bound, we train the Wae-derived model without low-resource constraints. We report monolingual (one language) and multilingual (all languages) versions of training a model on available data. We use the monolingual upper-bound EN model as a ''Translate-Test'' comparison. We also compare to monolingual and multilingual ''Translate-Train'' models to evaluate the value of gold samples compared to silver-standard training data. We follow previous work in using OPUS (Tiedemann, 2012) translations for MTOP and Google Translate (Wu et al., 2016) for MultiATIS ++SQL in all directions. Following Rosenbaum et al. (2022), we use a cross-lingual word alignment tool (SimAlign; Jalili Sabet et al., 2020) to project token positions from MTOP source to the parallel machine-translated output (e.g., to shift label `wordi` in EN to `wordj` in FR).

In all results, we report averages of five runs over different few-shot splits. For MTOP, we compare to ''silver-standard'' methods: ''Translate-and-Fill'' (Nicosia et al., 2021, TaF) which generates training data using MT, and CLASP (Rosenbaum et al., 2022) which uses MT and prompting to generate multilingual training data. We note that these models and dataset pre-processing methods are not public (we have confirmed that our methods are reasonably comparable with authors). For MultiATIS++SQL, we compare to XG-Reptile from (Sherborne and Lapata, 2023). This method uses meta-learning to approximate a ''task manifold'' using English data and constrain representations of target languages to be close to this manifold. This approach *implicitly* optimizes for cross-lingual transfer by regularizing the gradients for target languages to align with gradients for English. Minotaur differs in *explicitly* measuring the representation divergence across languages.

6 Results

We find that Minotaur validates our hypothesis that *explicitly* minimizing latent divergence

	EN	FR	ES	DE	HI	Avg.
Gold Monolingual	79.4	69.8	72.3	67.1	60.5	67.4 ± 5.3
Gold Multilingual	**81.3**	**75.7**	**77.2**	**72.8**	**71.6**	**74.4 ± 3.5**
Translate-Test	—	7.7	7.4	7.6	7.3	7.5 ± 0.2
Translate-Train Monolingual	—	41.7	31.4	50.1	32.2	38.9 ± 9.4
Translate-Train Multilingual	74.2	46.9	43.0	53.6	39.9	45.9 ± 5.9
Translate-Train Multilingual +MINOTAUR	77.5	59.9	60.2	61.6	42.2	56.0 ± 9.2
TaF mT5-large (Nicosia et al., 2021)	83.5	71.1	69.6	70.5	58.1	67.3 ± 6.2
TaF mT5-xxl (Nicosia et al., 2021)	**85.9**	**74.0**	71.5	**72.4**	61.9	70.0 ± 5.5
CLASP (Rosenbaum et al., 2022)	84.4	72.6	68.1	66.7	58.1	66.4 ± 6.1
MINOTAUR 1 SPIS	79.5 ± 0.4	71.9 ± 0.2	72.3 ± 0.1	68.4 ± 0.3	65.1 ± 0.1	69.4 ± 3.4
MINOTAUR 5 SPIS	77.7 ± 0.6	72.0 ± 0.6	73.6 ± 0.3	69.1 ± 0.5	68.2 ± 0.5	70.7 ± 2.5
MINOTAUR 10 SPIS	80.2 ± 0.4	72.8 ± 0.5	**74.9** ± 0.1	70.0 ± 0.7	**68.6** ± 0.5	**71.6** ± 2.8

Table 1: Accuracy on MTOP across (i) upper-bounds, (ii) translation baselines, (iii) ''silver-standard'' methods, and (iv) MINOTAUR with SPIS sampling at 1, 5 and 10. We report for *English*, *French*, *Spanish*, *German*, and *Hindi* with ± sample standard deviation. *Avg.* reports the target language average ± standard deviation across languages. Best result per-language and average for (i) and (ii)–(iv) are bolded.

		EN	FR	PT	ES	DE	ZH	Avg.
Gold Monolingual		72.3	73.0	71.8	67.2	73.4	**73.7**	71.9 ± 2.7
Gold Multilingual		**73.7**	**74.4**	**72.3**	**71.7**	**74.6**	71.3	**72.9** ± 1.5
Translate-Test		—	70.1	70.6	66.9	68.5	62.9	67.8 ± 3.1
Translate-Train Monolingual		—	62.2	53.0	65.9	55.4	67.1	60.8 ± 6.3
Translate-Train Multilingual		72.7	69.4	67.3	66.2	65.0	69.2	67.5 ± 1.9
Translate-Train Multilingual +MINOTAUR		74.8	73.7	71.3	68.5	70.1	69.0	70.6 ± 2.1
@1%	XG-REPTILE	73.8 ± 0.3	70.4 ± 1.8	70.8 ± 0.7	68.9 ± 2.3	69.1 ± 1.2	68.1 ± 1.2	69.5 ± 1.1
	MINOTAUR	**75.6** ± 0.4	**73.7** ± 0.6	**71.4** ± 0.9	**71.0** ± 0.5	**70.4** ± 1.3	**70.0** ± 0.9	**71.3** ± 1.4
@5%	XG-REPTILE	74.4 ± 1.3	73.0 ± 0.9	71.6 ± 1.1	**71.6** ± 0.7	71.1 ± 0.6	69.5 ± 0.5	71.4 ± 1.3
	MINOTAUR	**77.0** ± 1.0	**73.9** ± 1.4	**72.8** ± 1.1	71.1 ± 0.6	**72.8** ± 2.0	**72.3** ± 0.6	**72.6** ± 1.0
@10%	XG-REPTILE	75.8 ± 1.3	74.2 ± 0.2	72.8 ± 0.6	72.1 ± 0.7	73.0 ± 0.6	**72.8** ± 0.5	73.0 ± 0.8
	MINOTAUR	**79.8** ± 0.4	**75.6** ± 1.8	**75.4** ± 0.8	**73.2** ± 1.7	**76.8** ± 1.5	72.5 ± 0.7	**74.7** ± 1.8

Table 2: Denotation Accuracy on MultiATIS++SQL across (i) upper-bounds, (ii) translation baselines, and (iii) few-shot sampling for MINOTAUR compared to XG-REPTILE (Sherborne and Lapata, 2023) at 1%, 5%, and 10%. We report for *English*, *French*, *Portuguese*, *Spanish*, *German*, and *Chinese* ± sample standard deviation. *Avg.* reports the target language average ± standard deviation across languages. Best result per-language and average for (i) and (ii)–(iii) are bolded.

improves cross-lingual transfer with few training examples in the target language. As evidenced by our ablation studies, our technique is surprisingly robust and can function without any parallel data between languages. Overall, our method outperforms silver-standard data augmentation techniques (in Table 1) and few-shot meta-learning (in Table 2).

Cross-lingual Transfer in Task-Oriented Parsing Table 1 summarizes our results on MTOP against comparison models at multiple SPIS rates. Our system significantly improves on the ''Gold Monolingual'' upper-bound even at 1 SPIS by $> 2\%$ ($p < 0.01$, using a two-tailed sign test assumed hereafter). For few-shot transfer on MTOP, we observe strong cross-lingual transfer

even at 1 SPIS translating only 1.8% of the dataset. Few-shot transfer is competitive with a monolingual model using 100% of gold translated data and so represents a promising new strategy for this dataset. We note that even at a high SPIS rate of 100 (approximately $\sim 53.1\%$ of training data), MINOTAUR is significantly ($p < 0.01$) poorer than the ''Gold Multilingual'' upper-bound, highlighting that few-shot transfer is challenging on MTOP.

MINOTAUR outperforms all translation-based comparisons and augmenting ''Translate-Train Multilingual'' with our posterior alignment objective (+ MINOTAUR) yields a $+10.1\%$ average improvement. With equivalent data, this comparison shows that cross-lingual alignment by aligning each latent representation to the prior *only* (i.e., a WAE-based model) is weaker than cross-lingual alignment between posteriors.

Comparing to ''Silver-Standard'' Methods A more realistic comparison is between TaF (Nicosia et al., 2021) or CLASP (Rosenbaum et al., 2022), which optimize MT quality in their pipelines, and our method which uses sampled gold data. We outperform CLASP by >3% and TaF using mT5-large (Xue et al., 2021) by >2.1% at *all* sample rates. However, MINOTAUR requires > 5 SPIS sampling to improve upon TaF using mT5-xxl. We highlight that our model has only ~ 116 million parameters whereas CLASP uses AlexaTM-500M (FitzGerald et al., 2022) with 500 million parameters, mT5-large has 700 million parameters and mT5-xxl has 3.3 billion parameters. Relative to model size, our approach offers improved computational efficiency. The improvement of our method is mostly seen in languages typologically distant from English as MINOTAUR is always the strongest model for Hindi. In contrast, our method underperforms for English and German (more similar to EN) which may benefit from stronger pre-trained knowledge transfer within larger models. Our efficacy using gold data and a smaller model, compared to silver data in larger models, suggests a quality trade-off, constrained by computation, as a future study.

Cross-lingual Transfer in Executable Parsing The results for MultiATIS++SQL in Table 2 show similar trends. However, here MINOTAUR can outperform the upper-bounds, and sampling at $> 5\%$ significantly ($p < 0.01$) improves on ''Gold-Monolingual'' and is similar or better than ''Gold-Multilingual'' ($p < 0.05$). Further increasing the sample rate yields marginal gains. MINOTAUR generally improves on XG-REPTILE and performs on par *at a lower sample rate*, i.e., MINOTAUR at 1% sampling is closer to XG-REPTILE at 5% sampling. This suggests that our approach is *more sample efficient*, achieving greater accuracy with fewer samples. MINOTAUR requires < 10 epochs to train whereas XG-REPTILE reports ~ 50 training epochs, for poorer results.

$\mathbb{D}_{Z\|X}$	$\mathbb{D}_Z$	EN	FR	ES	DE	HI	Avg.
KL	—	78.3	70.6	73.1	67.0	66.6	69.3
W_2	—	78.6	72.1	74.3	68.7	67.4	70.6
—	MMD	78.7	72.3	74.3	68.8	67.5	70.7
KL	MMD	78.4	71.8	73.3	68.5	67.3	70.2
W_2	MMD	80.2	72.8	74.9	70.0	68.6	**71.6**

Table 3: Accuracy on MTOP at 10 SPIS permuting different alignment methods between individual-only ($\mathbb{D}_{Z|X}$), aggregate-only ($\mathbb{D}_Z$) and joint ($\mathbb{D}_{Z|X} + \mathbb{D}_Z$). The joint method using L_2-Wasserstein distance is empirically optimal but not significantly above the aggregate-only method ($p = 0.07$).

Despite demonstrating overall improvement, MINOTAUR is not universally superior. Notably, our performance on Chinese (ZH) is weaker than XG-REPTILE at 10% sampling and our method appears to benefit less from more data in comparison. The divergence minimization in MINOTAUR may be more functionally related to language similarity (dissimilar languages demanding greater distances to minimize) whereas the alignment via gradient constraints within meta-learning could be less sensitive to this phenomenon. These results, with the observation that MINOTAUR improves most on Hindi for MTOP, illustrate a need for more in-depth studies of cross-lingual transfer between *distant* and *lower resource* languages. Future work can consider more challenging benchmarks across a wider pool of languages (Ruder et al., 2023).

Contrasting Alignment Signals We report ablations of MINOTAUR on MTOP at 10 SPIS sampling. Table 3 considers each function for cross-lingual alignment outlined in Section 3.2 as an individual or composite element. The best approach, used in all other reported results, minimizes the Wasserstein distance W_2 for *individual*

	EN	FR	ES	DE	HI	Avg.
MMD	77.5	69.6	70.7	66.3	61.7	67.1
KL	77.9	69.8	70.9	66.5	62.1	67.3
L_2	77.1	69.2	70.3	65.8	61.7	66.8

Table 4: Accuracy on MTOP at 10 SPIS using non-parametric alignment without Z. Here the encoder output, $E_\phi(X)$ is input into decoder $G_\theta\left(E_\phi(X)\right)$. All approaches significantly underperform ($p < 0.01$) relative to Table 3.

divergence and MMD for *aggregate* divergence. W_2 is significantly superior to the Kullback-Leibler Divergence (KL) for minimizing *individual* posterior samples ($p < 0.01$ for individual and joint cases). The W_2 distance directly minimizes the Euclidean L_2 distance when variances of different languages are equivalent. This in turn is more similar to the Maximum Mean Discrepancy function (the best singular objective) which minimizes the distance between approximate ''means'' of each distribution i.e., between Z marginal distributions. Note that MMD and W_2 alignments are not significantly different ($p = 0.08$). The W_2 + MMD approach significantly outperforms all other combinations ($p < 0.01$). The identified strength of MMD, compared to methods for computing $\mathbb{D}_{Z|X}$, highlights that minimizing *aggregate* divergence is the main contributor to alignment with *individual* divergence as a weaker additional contribution.

Alignment without Latent Variables Table 4 considers alignment without the latent variable formulation on an encoder-decoder Transformer model (Vaswani et al., 2017). Here, the output of the encoder is not probabilistically bound without the parametric ''guidance'' of the Gaussian reparameterization. This is similar to analysis on explicit alignment from Wu and Dredze (2020). We test MMD, statistical KL divergence (e.g., $\sum_x p(x) log\left(\frac{p(x)}{q(x)}\right)$) and Euclidean L_2 distance as minimization functions and observe all techniques are significantly weaker ($p < 0.01$) than counterparts outlined in Table 3. This contrast suggests the smooth curvature and bounded structure of the Z parameterization contribute to effective cross-lingual alignment. Practically, these non-parametric approaches are challenging to implement. The lack of precise divergences (i.e., Equation (13) or Equation (12)) between representations leads to numerical underflow instability during training. This impeded alignment against reasonable comparisons such as cosine distance. Even using MMD, which does not require an exact solution, fared poorer without the bounding of the latent variable Z.

Alignment	EN	FR	ES	DE	HI	Avg.
Parallel Ref.	80.2	72.8	74.9	70.0	68.6	**71.6**
$\mathbb{D}_{Z\|X}$ only	78.9	67.3	68.3	64.6	59.4	64.9
$\mathbb{D}_Z$ only	77.6	71.5	72.9	68.4	67.2	**70.0**
$\mathbb{D}_{Z\|X} + \mathbb{D}_Z$	78.8	70.9	71.9	67.9	64.5	68.8

Table 5: Accuracy on MTOP at 10 SPIS using non-parallel inputs between languages in MINOTAUR. During training, we sample English input, x_{EN}, and an input in language l, x_l which is *not* a translation of x_{EN} for Equation (15). This approach weakens individual posterior alignment but identifies that MMD is the least sensitive to input parallelism.

Parallelism in Alignment We further investigate whether MINOTAUR induces cross-lingual transfer when aligning posterior samples from inputs which are *not* parallel (i.e., x_l is not a translation of x_{EN} and output LFs are not equivalent). We intuitively expect parallelism as necessary for the model to minimize divergence between representations with equivalent semantics.

As shown in Table 5, data parallelism is surprisingly *not required* using MMD to align marginal distributions *only*. The $\mathbb{D}_{Z|X}$ only and $\mathbb{D}_{Z|X} + \mathbb{D}_Z$ techniques significantly underperform relative to equivalent methods using parallel data ($p < 0.01$). This is largely expected because individual alignment between posterior samples which *should likely not be equivalent* could inject unnecessary noise into the learning process. However, MMD ($\mathbb{D}_Z$ only) is significantly ($p < 0.01$) above other methods with the closest performance to the parallel equivalent. This supports our interpretation that MMD aligns ''at the language level'' as minimization between languages should not mandate parallel data. For lower-resource scenarios, this approach could over-sample less data for cross-lingual transfer to the long tail of under-resourced languages.

Learning a Latent Semantic Structure We study the representation space learned from our

Model	Cosine (↑)	Top-1	Top-5	Top-10	MRR (↑)
MBART50	0.576	0.521	0.745	0.796	0.622
XG-REPTILE	0.844	0.797	0.949	0.963	0.865
MINOTAUR	**0.941**	**0.874**	**0.994**	**0.998**	**0.927**

Table 6: Average similarity between encodings of English and target languages for MultiATIS++SQL. Cosine similarity evaluates average distance between encodings of parallel sentences. Top-k evaluates if the parallel encoding is ranked within the k most cosine-similar vectors. Mean Reciprocal Rank (MRR) evaluates average position of parallel encodings ranked by similarity. Significant best results are bolded ($p < 0.01$).

method training on MultiATIS++SQL at 1% sampling for direct comparison to similar analysis from Sherborne and Lapata (2023). We compute sentence representations from the test set as the average of the $\mathbf{z}$ representations for each input utterance $\left(\frac{1}{T}\sum_i^T z_i\right)$. Table 6 compares between MINOTAUR, MBART50 (Tang et al., 2021) representations before training, and XG-REPTILE. The significant improvement in cross-lingual cosine similarity using MINOTAUR in Table 6 ($p < 0.01$) further supports how our proposed method learns improved cross-lingual similarity.

We also consider the most cosine-similar neighbors for each representation and test if the top-k closest representations are from a parallel utterance in a *different* language or some other utterance in the *same* language. Table 6 shows that $>$ 99% of representations learned by MINOTAUR have a parallel utterance within five closest representations and ∼50% improvement in mean-reciprocal ranking score (MRR) between parallel utterances. We interpret this as the representation space using MINOTAUR is more *semantically distributed* relative to MBART50, as representations for a given utterance are closer to semantic equivalents. We visualize this in Figure 3: The original pre-trained model has minimal cross-lingual overlap, whereas our system produces encodings with similarity aligned by *semantics* rather than *language*. MINOTAUR can rapidly adapt the pre-trained representations using an explicit alignment objective to produce a non-trivial informative latent structure. This formulation could have further utility within multilingual representation learning or information retrieval, e.g., to induce more coherent relationships between cross-lingual semantics.

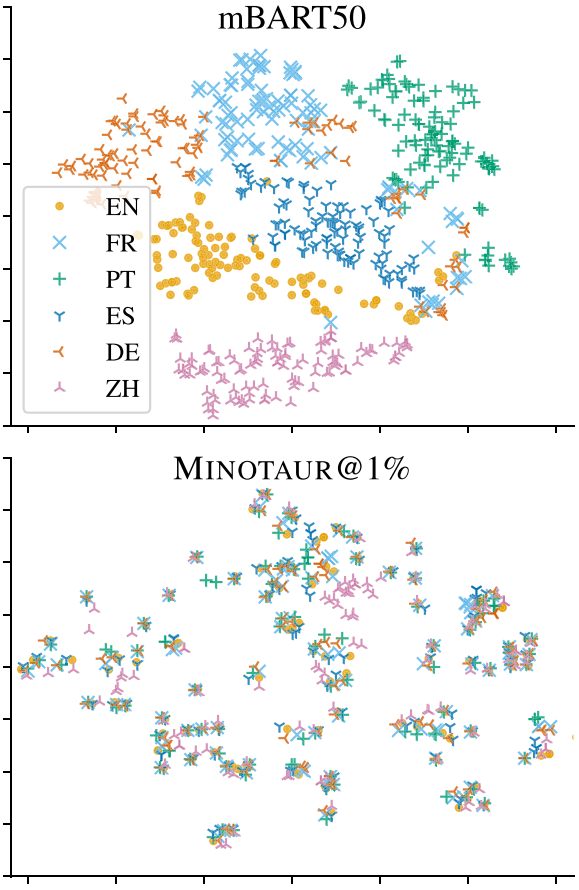

Figure 3: Visualization of MultiATIS++SQL encodings (test set; 25% random parallel sample) using t-SNE (van der Maaten and Hinton, 2008). Compared to MBART50, MINOTAUR organizes the latent space to be more *semantically distributed* across languages without monolingual separability.

Error Analysis We conduct an error analysis on MultiATIS++SQL examples correctly predicted by MINOTAUR and incorrectly predicted by baselines. The primary improvement arises from improved handling of multi-word expressions and language-specific modifiers. For example, adjectives in English are often multi-word adjectival phrases in French (e.g., ‘‘cheapest’’ → ‘‘le moins cher’’ or ‘‘earliest’’ → ‘‘à plus tot’’). Improved handling of this error type accounts for an average of 53% of improvement across languages with the highest in French (69%) and lowest in Chinese (38%). We hypothesize that a combination of aggregate and mean-pool individual alignment in MINOTAUR benefits this specific case where semantics are expressed in varying numbers of words

between languages. While this could be similarly approached using fine-grained token alignment labels, MINOTAUR improves transfer in this context without additional annotation. While this analysis is straightforward for French, it is unclear why the transfer to Chinese is weaker. A potential interpretation is that weaker transfer of multi-word expressions to Chinese could be related to poor tokenization. Sub-optimal sub-word tokenization of logographic or information-dense languages is an ongoing debate (Hofmann et al., 2022; Si et al., 2023) and exact explanations require further study. Translation-based models and weaker systems often generate malformed, non-executable SQL. Most additional improvement is due to a 23% boost in generating syntactically well-formed SQL evaluated within a database. Syntactic correctness is critical when a parser encounters a rare entity or unfamiliar linguistic construction and highlights how our model can better navigate inputs from languages minimally observed during training. This could potentially be further improved using recent incremental decoding advancements (Scholak et al., 2021).

7 Conclusion

We propose MINOTAUR, a method for few-shot cross-lingual semantic parsing leveraging Optimal Transport for knowledge transfer between languages. MINOTAUR uses a multi-level posterior alignment signal to enable sample-efficient semantic parsing of languages with few annotated examples. We identify how MINOTAUR aligns individual and aggregate representations to bootstrap parsing capability from English to multiple target languages. Our method is robust to different choices of alignment metrics and does not mandate parallel data for effective cross-lingual transfer. In addition, MINOTAUR learns more semantically distributed and language-agnostic latent representations with verifiably improved semantic similarity, indicating its potential application to improve cross-lingual generalization in a wide range of other tasks.

Acknowledgments

We thank the action editor and anonymous reviewers for their constructive feedback. The authors also thank Nikita Moghe, Mattia Opper, and N. Siddarth for their insightful comments on earlier versions of this paper. The authors (Sherborne, Lapata) gratefully acknowledge the support of the UK Engineering and Physical Sciences Research Council (grant EP/W002876/1). This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh (Hosking).

References

Sawsan Alqahtani, Garima Lalwani, Yi Zhang, Salvatore Romeo, and Saab Mansour. 2021. Using optimal transport as alignment objective for fine-tuning multilingual contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3904–3919, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.329

David Alvarez-Melis and Tommi Jaakkola. 2018. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1214

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, Seattle, Washington, USA. 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. https://doi.org/10.18653/v1/2020.emnlp-main.413

Jianpeng Cheng, Siva Reddy, Vijay Saraswat, and Mirella Lapata. 2019. Learning an executable

neural semantic parser. *Computational Linguistics*, 45(1):59–94. https://doi.org/10.1162/coli_a_00342

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2292–2300.

Jack G. M. FitzGerald, Shankar Ananthakrishnan, Konstantine Arkoudas, Davide Bernardi, Abhishek Bhagia, Claudio Delli Bovi, Jin Cao, Rakesh Chada, Amit Chauhan, Luoxin Chen, Anurag Dwarakanath, Satyam Dwivedi, Turan Gojayev, Karthik Gopalakrishnan, Thomas Gueudre, Dilek Hakkani-Tür, Wael Hamza, Jonathan Hueser, Kevin Martin Jose, Haidar Khan, Beiye Liu, Jianhua Lu, Alessandro Manzotti, Pradeep Natarajan, Karolina Owczarzak, Gokmen Oz, Enrico Palumbo, Charith Peris, Chandana Satya Prakash, Stephen Rawls, Andy Rosenbaum, Anjali Shenoy, Saleh Soltan, Mukund Harakere, Liz Tan, Fabian Triefenbach, Pan Wei, Haiyang Yu, Shuai Zheng, Gokhan Tur, and Prem Natarajan. 2022. Alexa teacher model: Pretraining and distilling multi-billion-parameter encoders for natural language understanding systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, pages 2893–2902, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3534678.3539173

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-2501

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.

Milan Gritta, Ruoyu Hu, and Ignacio Iacobacci. 2022. CrossAligner & co: Zero-shot transfer methods for task-oriented cross-lingual natural language understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4048–4061, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.319

Yingmei Guo, Linjun Shou, Jian Pei, Ming Gong, Mingxing Xu, Zhiyong Wu, and Daxin Jiang. 2021. Learning from multiple noisy augmented data sets for better cross-lingual spoken language understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3226–3237, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.259

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24–27, 1990*. https://doi.org/10.3115/116580.116613

Daniel Hershcovich, Zohar Aizenbud, Leshem Choshen, Elior Sulem, Ari Rappoport, and Omri Abend. 2019. SemEval-2019 task 1: Cross-lingual semantic parsing with UCCA. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1–10, Minneapolis, Minnesota, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-2001

Valentin Hofmann, Hinrich Schuetze, and Janet Pierrehumbert. 2022. An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–393, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-short.43

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 *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13–18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Zhiqi Huang, Puxuan Yu, and James Allan. 2023. Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, pages 1048–1056, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3539597.3570468

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.147

Zhanming Jie and Wei Lu. 2014. Multilingual semantic parsing : Parsing multiple languages into semantic representations. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1291–1301, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Bevan Jones, Mark Johnson, and Sharon Goldwater. 2012. Semantic parsing with Bayesian tree transducers. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 488–496, Jeju Island, Korea. Association for Computational Linguistics.

Aishwarya Kamath and Rajarshi Das. 2019. A survey on semantic parsing. In *Proceedings of the 1st Conference on Automated Knowledge Base Construction, AKBC*. Amherst, MA, USA.

Lev Kantorovich. 1958. On the translocation of masses. *Management Science*, 5(1):1–4. https://doi.org/10.1287/mnsc.5.1.1

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*. https://doi.org/10.48550/arXiv.1412.6980

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings*. https://doi.org/10.48550/arXiv.1312.6114

Tomáš Kočiský, Gábor Melis, Edward Grefenstette, Chris Dyer, Wang Ling, Phil Blunsom, and Karl Moritz Hermann. 2016. Semantic parsing with semi-supervised sequential autoencoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1078–1087, Austin, Texas. Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1116

Thomas Kollar, Danielle Berry, Lauren Stuart, Karolina Owczarzak, Tagyoung Chung, Lambert Mathias, Michael Kayser, Bradford Snow, and Spyros Matsoukas. 2018. The Alexa meaning representation language. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 177–184, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-3022

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-2012

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. https://doi.org/10.18653/v1/2021.eacl-main.257

Haoyang Li, Jing Zhang, Cuiping Li, and Hong Chen. 2023. Resdsql: Decoupling schema linking and skeleton parsing for text-to-sql. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13067–13075. https://doi.org/10.1609/aaai.v37i11.26535

Percy Liang. 2016. Learning executable semantic parsers for natural language understanding. *Communications of the ACM*, 59(9):68–76. https://doi.org/10.1145/2866568

Shining Liang, Linjun Shou, Jian Pei, Ming Gong, Wanli Zuo, Xianglin Zuo, and Daxin Jiang. 2022. Label-aware multi-level contrastive learning for cross-lingual spoken language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9903–9918, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.673

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 dataset for 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. https://doi.org/10.18653/v1/2020.emnlp-main.484

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1500

Kelly Marchisio, Ali Saad-Eldin, Kevin Duh, Carey Priebe, and Philipp Koehn. 2022. Bilingual lexicon induction for low-resource languages using graph matching via optimal transport. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2545–2561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.164

Emile Mathieu, Tom Rainforth, N. Siddharth, and Yee Whye Teh. 2019. Disentangling disentanglement in variational autoencoders. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 4402–4412. PMLR.

Gaspard Monge. 1781. Mémoire sur la théorie des déblais et des remblais. *Mémoires de mathématique et de physique, presentés à l'Académie royale des sciences, par divers sçavans & lûs dans ses assemblées*, pages 666–704.

Thong Thanh Nguyen and Anh Tuan Luu. 2022. Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 – March 1, 2022*, pages 11103–11111. AAAI Press. https://doi.org/10.1609/aaai.v36i10.21359

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. https://doi.org/10.18653/v1/2021.findings-emnlp.279

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory

Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.255

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. https://doi.org/10.18653/v1/P19-1493

Bowen Qin, Binyuan Hui, Lihan Wang, Min Yang, Jinyang Li, Binhua Li, Ruiying Geng, Rongyu Cao, Jian Sun, Luo Si, Fei Huang, and Yongbin Li. 2022a. A survey on text-to-SQL parsing: Concepts, methods, and future directions. *ArXiv preprint*, abs/2208.13629.

Libo Qin, Qiguang Chen, Tianbao Xie, Qixin Li, Jian-Guang Lou, Wanxiang Che, and Min-Yen Kan. 2022b. GL-CLeF: A global–local contrastive learning framework for cross-lingual spoken language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2677–2686, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.191

Karthik Raman, Iftekhar Naim, Jiecao Chen, Kazuma Hashimoto, Kiran Yalasangi, and Krishna Srinivasan. 2022. Transforming sequence tagging into a Seq2Seq task. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11856–11874, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.813

Siva Reddy, Oscar Täckström, Slav Petrov, Mark Steedman, and Mirella Lapata. 2017. Universal semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 89–101, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1009

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21–26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1278–1286. JMLR.org.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Marco Damonte, Isabel Groves, and Amir Saffari. 2022. CLASP: Few-shot cross-lingual data augmentation for semantic parsing. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 444–462, Online only. Association for Computational Linguistics.

Sebastian Ruder, Jonathan H. Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel A. Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana L. Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David I. Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. 2023. Xtreme-up: A user-centric scarce-data benchmark for under-represented languages. https://doi.org/10.48550/arXiv.2305.11938

Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. 2021. PICARD: Parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9895–9901, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.779

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. https://doi.org/10.18653/v1/2022.acl-long.285

Tom Sherborne and Mirella Lapata. 2023. Meta-learning a cross-lingual manifold for semantic parsing. *Transactions of the Association for Computational Linguistics*, 11:49–67. https://doi.org/10.1162/tacl_a_00533

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. https://doi.org/10.18653/v1/2020.findings-emnlp.45

Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. 2022. XRICL: Cross-lingual retrieval-augmented in-context learning for cross-lingual text-to-SQL semantic parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5248–5259, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.384

Chenglei Si, Zhengyan Zhang, Yingfa Chen, Fanchao Qi, Xiaozhi Wang, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2023. Sub-character tokenization for Chinese pretrained language models. *Transactions of the Association for Computational Linguistics*, 11:469–487. https://doi.org/10.1162/tacl_a_00560

Raymond Hendy Susanto and Wei Lu. 2017a. Neural architectures for multilingual semantic parsing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 38–44, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2007

Raymond Hendy Susanto and Wei Lu. 2017b. Semantic parsing with neural hybrid trees. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4–9, 2017, San Francisco, California, USA*, pages 3309–3315. AAAI Press.

Asuka Takatsu. 2011. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.304

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ilya O. Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. 2018. Wasserstein auto-encoders. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 – May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497,

Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.197

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Cedric Villani. 2008. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-71050-9

Prince Zizhuang Wang and William Yang Wang. 2019. Riemannian normalizing flow on variational Wasserstein autoencoder for text modeling. 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 284–294, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1025

Zhiruo Wang, Grace Cuenca, Shuyan Zhou, Frank F. Xu, and Graham Neubig. 2023. MCoNaLa: A benchmark for code generation from multiple natural languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 265–273, Dubrovnik, Croatia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-eacl.20

John Wieting, Jonathan H. Clark, William W. Cohen, Graham Neubig, and Taylor Berg-Kirkpatrick. 2023. Beyond contrastive learning: A variational generative model for multilingual retrieval. https://doi.org/10.18653/v1/2023.acl-long.673

Shijie Wu and Mark Dredze. 2020. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv preprint*, abs/1609.08144. https://doi.org/10.48550/arXiv.1609.08144

Menglin Xia and Emilio Monti. 2021. Multilingual neural semantic parsing for low-resourced languages. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 185–194, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.starsem-1.17

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. https://doi.org/10.18653/v1/2020.emnlp-main.410

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. https://doi.org/10.18653/v1/2021.naacl-main.41

Pengcheng Yin, Chunting Zhou, Junxian He, and Graham Neubig. 2018. StructVAE: Tree-structured latent variable models for semi-supervised semantic parsing. In *Proceedings of*

the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 754–765, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1070

John M. Zelle and Raymond J. Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the 13th National Conference on Artificial Intelligence - Volume 2*, AAAI'96, pages 1050–1055.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. 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 5751–5767, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.447