Neural Architectures for Multilingual Semantic Parsing

Raymond Hendy Susanto and **Wei Lu**
Singapore University of Technology and Design
{raymond_susanto,luwei}@sutd.edu.sg

Abstract

In this paper, we address semantic parsing in a multilingual context. We train one multilingual model that is capable of parsing natural language sentences from multiple different languages into their corresponding formal semantic representations. We extend an existing sequence-to-tree model to a multi-task learning framework which shares the decoder for generating semantic representations. We report evaluation results on the multilingual GeoQuery corpus and introduce a new multilingual version of the ATIS corpus.

1 Introduction

In this work, we address *multilingual* semantic parsing – the task of mapping natural language sentences coming from multiple different languages into their corresponding formal semantic representations. We consider two multilingual scenarios: 1) the *single-source* setting, where the input consists of a single sentence in a single language, and 2) the *multi-source* setting, where the input consists of parallel sentences in multiple languages. Previous work handled the former by means of monolingual models (Wong and Mooney, 2006; Lu et al., 2008; Jones et al., 2012), while the latter has only been explored by Jie and Lu (2014) who ensembled many monolingual models together. Unfortunately, training a model for each language separately ignores the shared information among the source languages, which may be potentially beneficial for typologically related languages. Practically, it is also inconvenient to train, tune, and configure a new model for each language, which can be a laborious process.

In this work, we propose a parsing architecture that accepts as input sentences in several languages. We extend an existing sequence-to-tree model (Dong and Lapata, 2016) to a multi-task learning framework, motivated by its success in other fields, e.g., neural machine translation (MT) (Dong et al., 2015; Firat et al., 2016). Our model consists of *multiple encoders*, one for each language, and *one decoder* that is shared across source languages for generating semantic representations. In this way, the proposed model potentially benefits from having a generic decoder that works well across languages. Intuitively, the model encourages each source language encoder to find a common structured representation for the decoder. We further modify the attention mechanism (Bahdanau et al., 2015) to integrate multi-source information, such that it can learn where to focus during parsing; i.e., which input positions in which languages.

Our contributions are as follows:

- We investigate semantic parsing in two multilingual scenarios that are relatively unexplored in past research,
- We present novel extensions to the sequence-to-tree architecture that integrates multilingual information for semantic parsing, and
- We release a new ATIS semantic dataset annotated in two new languages.

2 Related Work

In this section, we summarize semantic parsing approaches from previous works. Wong and Mooney (2006) created WASP, a semantic parser based on statistical machine translation. Lu et al. (2008) proposed generative hybrid tree structures, which were augmented with a discriminative reranker. CCG-based semantic parsing systems have been developed, such as ZC07 (Zettlemoyer and Collins, 2007) and UBL (Kwiatkowski et al.,

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 38–44
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-2007

2010). Researchers have proposed sequence-to-sequence parsing models (Jia and Liang, 2016; Dong and Lapata, 2016; Kočiský et al., 2016). Recently, Susanto and Lu (2017) extended the hybrid tree with neural features.

Recent progress in multilingual NLP has moved towards building a unified model that can work across different languages, such as in multilingual dependency parsing (Ammar et al., 2016), multilingual MT (Firat et al., 2016), and multilingual word embedding (Guo et al., 2016). Nonetheless, multilingual approaches for semantic parsing are relatively unexplored, which motivates this work. Jones et al. (2012) evaluated an individually-trained tree transducer on a multilingual semantic dataset. Jie and Lu (2014) ensembled monolingual hybrid tree models on the same dataset.

3 Model

In this section, we describe our approach to multilingual semantic parsing, which extends the sequence-to-tree model by Dong and Lapata (2016). Unlike the mainstream approach that trains one monolingual parser per source language, our approach integrates N *encoders*, one for each language, into a single model. This model encodes a sentence from the n-th language $X = x_1, x_2, ..., x_{|X|}$ as a vector and then uses a shared *decoder* to decode the encoded vector into its corresponding logical form $Y = y_1, y_2, ..., y_{|Y|}$. We consider two types of input: 1) a single sentence in one of N languages in the *single-source* setting and 2) parallel sentences in N languages in the *multi-source* setting. We elaborate on each setting in Section 3.1 and 3.2, respectively.

The encoder is implemented as a unidirectional RNN with long short-term memory (LSTM) units (Hochreiter and Schmidhuber, 1997), which takes a sequence of natural language tokens as input. Similar to previous multi-task frameworks, e.g., in neural MT (Firat et al., 2016; Zoph and Knight, 2016), we create one encoder per source language, i.e., $\{\Psi^n_{\text{enc}}\}^N_{n=1}$. For the n-th language, it updates the hidden vector at time step t by:

$$\mathbf{h}^n_t = \Psi^n_{\text{enc}}(\mathbf{h}^n_{t-1}, \mathbf{E}^n_x[x_t]) \qquad (1)$$

where Ψ^n_{enc} is the LSTM function and $\mathbf{E}^n_x \in \mathbb{R}^{|V|\times d}$ is an embedding matrix containing row vectors of the source tokens in the n-th language. Each encoder may be configured differently, such as by the number of hidden units and the embedding dimension for the source symbol.

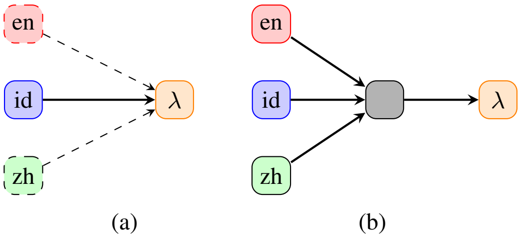

Figure 1: Illustration of the model with three language encoders and a shared logical form decoder (in λ-calculus). Two scenarios are considered: (a) *single-source* and (b) *multi-source* with a combiner module (in grey color).

In the basic sequence-to-sequence model, the decoder generates each target token in a linear fashion. However, in semantic parsing, such a model ignores the hierarchical structure of logical forms. In order to alleviate this issue, Dong and Lapata (2016) proposed a decoder that generates logical forms in a top-down manner, where they define a "non-terminal" token <n> to indicate subtrees. At each depth in the tree, logical forms are generated sequentially until the end-of-sequence token is output.

Unlike in the single language setting, here we define a single, shared decoder Ψ_{dec} as opposed to one decoder per source language. We augment the parent non-terminal's information $\mathbf{p}$ when computing the decoder state $\mathbf{z}_t$, as follows:

$$\mathbf{z}_t = \Psi_{\text{dec}}(\mathbf{z}_{t-1}, \mathbf{E}_y[\tilde{y}_{t-1}], \mathbf{p}) \qquad (2)$$

where Ψ_{dec} is the LSTM function and $\tilde{y}_{t-1}$ is the previous target symbol.

The attention mechanism (Bahdanau et al., 2015; Luong et al., 2015) computes a time-dependent context vector $\mathbf{c}_t$ (as defined later in Section 3.1 and 3.2), which is subsequently used for computing the probability distribution over the next symbol, as follows:

$$\tilde{\mathbf{z}}_t = \tanh(\mathbf{U}\mathbf{z}_t + \mathbf{V}\mathbf{c}_t) \qquad (3)$$
$$p(y_t|y_{<t}, X) \propto \exp(\mathbf{W}\tilde{\mathbf{z}}_t) \qquad (4)$$

where $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{W}$ are weight matrices. Finally, the model is trained to maximize the following conditional log-likelihood:

$$\mathcal{L}(\theta) = \sum_{(X,Y)\in\mathcal{D}} \sum_{t=1}^{|Y|} \log p(y_t|y_{<t}, X) \qquad (5)$$

where (X, Y) refers to a ground-truth sentence-semantics pair in the training data $\mathcal{D}$.

We use the same formulation above for the encoders and the decoder in both multilingual settings. Each setting differs in terms of: 1) the decoder state initialization, 2) the computation of the context vector $\mathbf{c}_t$, and 3) the training procedure, which are described in the following sections.

3.1 Single-Source Setting

In this setting, the input is a source sentence coming from the n-th language. Figure 1 (a) depicts a scenario where the model is parsing Indonesian input, with English and Chinese being non-active.

The last state of the n-th encoder is used to initialize the first state of the decoder. We may need to first project the encoder vector into a suitable dimension for the decoder, i.e., $\mathbf{z}_0 = \phi^n_{\text{dec}}(\mathbf{h}^n_{|X|})$, where ϕ^n_{dec} can be an affine transformation. Similarly, we may do so before computing the attention scores, i.e., $\tilde{\mathbf{h}}^n_k = \phi^n_{\text{att}}(\mathbf{h}^n_k)$. Then, we compute the context vector $\mathbf{c}^n_t$ as a weighted sum of the hidden vectors in the n-th encoder:

$$\alpha^n_{k,t} = \frac{\exp(\tilde{\mathbf{h}}^n_k \cdot \mathbf{z}_t)}{\sum_{k'=1}^{|X|} \exp(\tilde{\mathbf{h}}^n_{k'} \cdot \mathbf{z}_t)} \tag{6}$$

$$\mathbf{c}^n_t = \sum_{k=1}^{|X|} \alpha^n_{k,t} \tilde{\mathbf{h}}^n_k \tag{7}$$

We set $\mathbf{c}_t = \mathbf{c}^n_t$ for computing Equation 3. We propose two variants of the model under this setting. In the first version, we define separate weight matrices for each language, i.e., $\{\mathbf{U}^n, \mathbf{V}^n, \mathbf{W}^n\}_{n=1}^N$. In the second version, the three weight matrices are shared across languages, essentially reducing the number of parameters by a factor of N.

The training data consists of the union of sentence-semantics pairs in N languages, where the source sentences are not necessarily parallel. We implement a scheduling mechanism that cycles through all languages during training, one language at a time. Specifically, model parameters are updated after one batch from one language before moving to the next one. Similar to Firat et al. (2016), this mechanism prevents excessive updates from a specific language.

3.2 Multi-Source Setting

In this setting, the input are semantically equivalent sentences in N languages. Figure 1 (b) depicts a scenario where the model is parsing English, Indonesian, and Chinese *simultaneously*. It includes a *combiner* module (denoted by the grey box), which we will explain next.

The decoder state at the first time step is initialized by first combining the N final states from each encoder, i.e., $\mathbf{z}_0 = \phi_{\text{init}}(\mathbf{h}^1_{|X|}, \cdots, \mathbf{h}^N_{|X|})$, where we implement ϕ_{init} by max-pooling.

We propose two ways of computing $\mathbf{c}_t$ that integrates source-side information from multiple encoders. First, we consider **word-level combination**, where we combine N encoder states at every time step, as follows:

$$\alpha^n_{k,t} = \frac{\exp(\tilde{\mathbf{h}}^n_k \cdot \mathbf{z}_t)}{\sum_{n'=1}^{N} \sum_{k'=1}^{|X|} \exp(\tilde{\mathbf{h}}^{n'}_{k'} \cdot \mathbf{z}_t)} \tag{8}$$

$$\mathbf{c}_t = \sum_{n=1}^{N} \sum_{k=1}^{|X|} \alpha^n_{k,t} \tilde{\mathbf{h}}^n_k \tag{9}$$

Alternatively, in **sentence-level combination**, we first compute the context vector for each language in the same way as Equation 6 and 7. Then, we perform a simple concatenation of N context vectors: $\mathbf{c}_t = \left[\mathbf{c}^1_t; \cdots; \mathbf{c}^N_t\right]$.

Unlike the single-source setting, the training data consists of N-way parallel sentence-semantics pairs. That is, each training instance consists of N semantically equivalent sentences and their corresponding logical form.

4 Experiments and Results

4.1 Datasets and Settings

We conduct our experiments on two multilingual benchmark datasets, which we describe below. Both datasets use a meaning representation based on lambda calculus.

The GeoQuery (GEO) dataset is a standard benchmark evaluation for semantic parsing. The multilingual version consists of 880 instances of natural language queries related to US geography facts in four languages (English, German, Greek, and Thai) (Jones et al., 2012). We use the standard split which consists of 600 training examples and 280 test examples.

The ATIS dataset contains natural language queries to a flight database. The data is split into 4,434 instances for training, 491 for development, and 448 for evaluation, same as Zettlemoyer and Collins (2007). The original version only includes English. In this work, we annotate the corpus in Indonesian and Chinese. The Chinese corpus was annotated (with segmentations) by hiring professional translation service. The Indonesian corpus was annotated by a native Indonesian speaker.

We use the same pre-processing as Dong and Lapata (2016), where entities and numbers are replaced with their type names and unique IDs.[1] English words are stemmed using NLTK (Bird et al., 2009). Each query is paired with its corresponding semantic representation in lambda calculus (Zettlemoyer and Collins, 2005).

In all experiments, following Dong and Lapata (2016), we use a one-layer LSTM with 200-dimensional cells and embeddings. We use a minibatch size of 20 with RMSProp updates (Tieleman and Hinton, 2012) for a fixed number of epochs, with gradient clipping at 5. Parameters are uniformly initialized at [-0.08,0.08] and regularized using dropout (Srivastava et al., 2014). Input sequences are reversed. See Appendix A for detailed experimental settings.

For each model configuration, all experiments are repeated 3 times with different random seed values, in order to make sure that our findings are reliable. We found empirically that the random seed may affect Seq2Tree performance. This is especially important due to the relatively small dataset. As previously done in multitask sequence-to-sequence learning (Luong et al., 2016), we report the average performance for the baseline and our model. The evaluation metric is defined in terms of exact match accuracy with the ground-truth logical forms. See Appendix B for the accuracy of individual runs.

4.2 Results

Table 1 compares the performance of the monolingual sequence-to-tree model (Dong and Lapata, 2016), Single, and our multilingual model, Multi, with separate and shared output parameters under the single-source setting as described in Section 3.1. On average, both variants of the multilingual model outperform the monolingual model by up to 1.34% average accuracy on Geo. Parameter sharing is shown to be helpful, in particular for Geo. We observe that the average performance increase on Atis mainly comes from Chinese and Indonesian. We also learn that although including English is often helpful for the other languages, it may affect its individual performance.

Table 2 shows the average performance on multi-source parsing by combining 3 to 4 languages for Geo and 2 to 3 languages for ATIS. For Ranking, we combine the predictions from each language by selecting the one with the highest probability. Indeed, we observe that system combination at the *model* level is able to give better performance on average (up to 4.29% on Geo) than doing so at the *output* level. Combining at the word level and sentence level shows comparable performance on both datasets. It can be seen that the benefit is more apparent when we include English in the system combination.

	Single	Multi	
		separate	shared
Geo			
en	84.40	85.00	**85.48**
de	70.24	71.19	**72.86**
el	74.40	75.12	**75.60**
th	72.86	72.26	**73.33**
avg.	75.48	75.89	**76.82**
Atis			
en	**81.85**	81.40	81.77
id	74.85	74.03	**75.45**
zh	73.66	**75.89**	73.96
avg.	76.79	**77.11**	77.06

Table 1: Single-source parsing results in terms of average accuracy % over 3 runs. Best results are in **bold**.

	Ranking	Multi	
		word	sentence
Geo			
en+de+el	83.21	85.48	**86.43**
en+de+th	82.02	**86.19**	85.48
en+el+th	82.62	**85.60**	85.24
de+el+th	**79.64**	72.14	76.43
en+de+el+th	82.50	85.48	**86.79**
Atis			
en+id	82.81	**83.93**	83.78
en+zh	82.81	**82.96**	**82.96**
id+zh	**78.50**	76.79	77.75
en+id+zh	83.11	82.22	**83.85**

Table 2: Multi-source parsing results in terms of average accuracy % over 3 runs. Best results are in **bold**.

Regarding comparison to previous monolingual works, we want to highlight that there exist two different versions of the GeoQuery dataset annotated with completely different semantic representations: semantic tree and lambda calculus. As noted in Section 5 of Lu (2014), results obtained from these two versions are not comparable. We use lambda calculus same as Dong and Lapata (2016). Under the multilingual setting, the closest work is Jie and Lu (2014). Nonetheless, they used the semantic tree version of GeoQuery. They eval-

[1] See Section 3.6 of (Dong and Lapata, 2016).

Model	**Input**	**Output**
SINGLE (en)	list the airlines with flights to or from ci0	lambda $0 e (and (airline $0) (exists $1 (and (flight $1) (or (from $1 ci0) (to $1 ci0)) (airline $1 $0))))
SINGLE (id)	daftarkan maskapai dengan penerbangan ke atau dari ci0	lambda $0 e (and (airline $0) (exists $1 (and (flight $1) (from $1 ci0) (airline $1 $0))))
SINGLE (zh)	请 列出 有 航班 起降 ci0 的 航空 公司	lambda $0 e (and (airline $0) (services $0 ci0))
MULTI	(en+id+zh)	lambda $0 e (exists $1 (and (flight $1) (or (from $1 ci0) (to $1 ci0)) (= (airline:e $1) $0)))
GOLD	(en+id+zh)	lambda $0 e (exists $1 (and (flight $1) (or (from $1 ci0) (to $1 ci0)) (= (airline:e $1) $0)))

Table 3: Example output from monolingual and multilingual models trained on ATIS.

Model	Number of parameters	
	GEO	ATIS
SINGLE/RANKING	3.7×10^6	3.1×10^6
MULTI (single)		
- separate	2.3×10^6	2.1×10^6
- shared	2.0×10^6	1.9×10^6
MULTI (multi)		
- word	2.0×10^6	1.9×10^6
- sentence	2.1×10^6	1.9×10^6

Table 4: Model size

uated extrinsically on a database query task while we use exact match accuracy, so their work is not directly comparable to ours.

5 Analysis

In this section, we report a qualitative analysis of our multilingual model. Table 3 shows example output from the monolingual model, SINGLE, trained on the three languages in ATIS and the multilingual model, MULTI, with sentence-level combination. This example demonstrates a scenario when the multilingual model successfully parses the three input sentences into the correct logical form, whereas the individual models are unable to do so.

Figure 2 shows the alignments produced by MULTI (sentence) when parsing ATIS in the multi-source setting. Each cell in the alignment matrix corresponds to $\alpha^n_{k,t}$ which is computed by Equation 6. Semantically related words are strongly aligned, such as the alignments between *ground* (en), *darat* (id), 地面 (zh) and *ground_transport*. This shows that such correspondences can be jointly learned by our multilingual model.

In Table 4, we summarize the number of parameters in the baseline and our multilingual model. The number of parameters in SINGLE and RANKING is equal to the sum of the number of parameters in their monolingual components. It can be seen that the size of our multilingual model is about 50-60% smaller than that of the baseline.

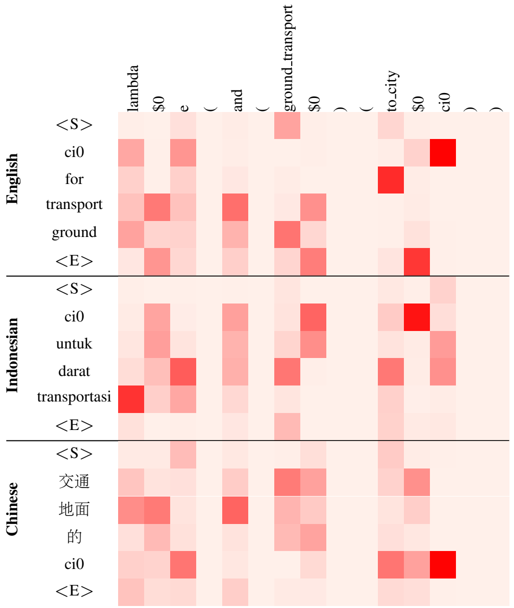

Figure 2: Attention score matrices computed by MULTI when parsing English, Indonesian, and Chinese inputs from ATIS. Darker color represents higher attention score.

6 Conclusion

We have presented a multilingual semantic parser that extends the sequence-to-tree model to a multi-task learning framework. Through experiments, we show that our multilingual model performs better on average than 1) monolingual models in the single-source setting and 2) ensemble ranking in the multi-source setting. We hope that this work will stimulate further research in multilingual semantic parsing. Our code and data is available at http://statnlp.org/research/sp/.

Acknowledgments

We would like to thank Christopher Bryant, Li Dong, and the anonymous reviewers for helpful comments and feedback. This work is supported by MOE Tier 1 grant SUTDT12015008, and is partially supported by project 61472191 under the National Natural Science Foundation of China.

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics* 4:431–444.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of ACL*. https://doi.org/10.3115/v1/P15-1166.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. In *Proceedings of ACL*. https://doi.org/10.18653/v1/P16-1004.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of NAACL*. https://doi.org/10.18653/v1/N16-1101.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A representation learning framework for multi-source transfer parsing. In *Proceedings of AAAI*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Robin Jia and Percy Liang. 2016. Data recombination for neural semantic parsing. In *Proceedings of ACL*. https://doi.org/10.18653/v1/P16-1002.

Zhanming Jie and Wei Lu. 2014. Multilingual semantic parsing: Parsing multiple languages into semantic representations. In *Proceedings of COLING*. http://aclweb.org/anthology/C14-1122.

Bevan Keeley Jones, Mark Johnson, and Sharon Goldwater. 2012. Semantic parsing with bayesian tree transducers. In *Proceedings of ACL*. http://aclweb.org/anthology/P12-1051.

Tomáš Kočiskỳ, 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 EMNLP*. https://doi.org/10.18653/v1/D16-1116.

Tom Kwiatkowski, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2010. Inducing probabilistic ccg grammars from logical form with higher-order unification. In *Proceedings of EMNLP*. http://aclweb.org/anthology/D10-1119.

Wei Lu. 2014. Semantic parsing with relaxed hybrid trees. In *Proceedings of EMNLP*. http://aclweb.org/anthology/D14-1137.

Wei Lu, Hwee Tou Ng, Wee Sun Lee, and Luke S Zettlemoyer. 2008. A generative model for parsing natural language to meaning representations. In *Proceedings of EMNLP*. http://aclweb.org/anthology/D08-1082.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *Proceedings of ICLR*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP*. https://doi.org/10.18653/v1/D15-1166.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958.

Raymond Hendy Susanto and Wei Lu. 2017. Semantic parsing with neural hybrid trees. In *Proceedings of AAAI*.

Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* 4(2).

Yuk Wah Wong and Raymond J Mooney. 2006. Learning for semantic parsing with statistical machine translation. In *Proceedings of NAACL*. http://aclweb.org/anthology/N06-1056.

Luke S Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of UAI*.

Luke S Zettlemoyer and Michael Collins. 2007. Online learning of relaxed ccg grammars for parsing to logical form. In *Proceedings of EMNLP-CoNLL*. http://aclweb.org/anthology/D07-1071.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of NAACL*. https://doi.org/10.18653/v1/N16-1004.

A Hyperparameters

Table 5 lists the number of training epochs and the dropout probability used in the LSTM cell and the hidden layers before the softmax classifiers, which were chosen based on preliminary experiments on a held-out dataset. We use a training schedule where we switch to the next language after training one mini-batch for GEO and 500 for ATIS. For

	Single			Multi					
				separate			shared		
	1	2	3	1	2	3	1	2	3
Geo									
en	**87.14**	83.57	82.50	85.71	83.93	85.36	85.36	83.93	**87.14**
de	70.00	70.36	70.36	71.79	71.79	70.00	73.57	**73.93**	71.07
el	76.43	72.50	74.29	**77.14**	72.14	76.07	76.43	74.64	75.71
th	72.50	73.57	72.50	72.14	72.14	72.50	72.50	71.07	**76.43**
Atis									
en	**84.60**	79.24	81.70	82.14	81.03	81.03	82.59	80.36	82.37
id	75.67	74.55	74.33	75.67	72.54	73.88	**76.56**	75.45	74.33
zh	74.33	73.66	72.99	74.11	76.12	**77.46**	75.67	72.54	73.66

Table 6: Single-source parsing results showing the accuracy of the 3 runs. Best results are in **bold**.

	Ranking			Multi					
				word			sentence		
	1	2	3	1	2	3	1	2	3
Geo									
en+de+el	85.00	82.50	82.14	87.14	84.64	84.64	**87.50**	85.36	86.43
en+de+th	84.29	81.07	80.71	**87.86**	85.00	85.71	85.71	86.43	84.29
en+el+th	84.29	82.14	81.43	**87.50**	84.29	85.00	84.64	85.71	85.36
de+el+th	**80.00**	79.29	79.64	71.07	72.86	72.50	77.86	74.64	76.79
en+de+el+th	83.93	81.79	81.79	85.71	86.07	84.64	**87.50**	86.79	86.07
Atis									
en+id	83.48	82.14	82.81	83.48	83.48	84.82	85.27	80.58	**85.49**
en+zh	84.60	80.80	83.04	83.26	82.14	83.48	**85.49**	80.13	83.26
id+zh	79.24	78.57	77.68	77.46	78.35	74.55	**80.58**	78.13	74.55
en+id+zh	84.15	81.92	83.26	82.14	81.25	83.26	**85.49**	81.03	85.04

Table 7: Multi-source parsing results showing the accuracy of the 3 runs. Best results are in **bold**.

all multilingual models, we initialize the encoders using the encoder weights learned by the monolingual models. For the multi-source setting, we also initialize the decoder using the first language in the list of the combined languages.

B Additional Experimental Results

In Table 6 and 7, we report the accuracy of the 3 runs for each model and dataset. In both settings, we observe that the best accuracy on both datasets is often achieved by Multi. This is the same conclusion that we reached when averaging the results over all runs.

	#epochs	dropout (LSTM)	dropout (output layer)
Geo			
Single	90	0.1	0.4
Multi (single)	340	0.1	0.4
Multi (multi)	150	0.1	0.4
Atis			
Single	130	0.3	0.3
Multi (single)	390	0.3	0.3
Multi (multi)	250	0.3	0.3

Table 5: Hyperparameter values