Frustratingly Simple but Surprisingly Strong: Using Language-Independent Features for Zero-shot Cross-lingual Semantic Parsing

Jingfeng Yang[‡] **Federico Fancellu**[†] **Bonnie Webber**[◇] **Diyi Yang**[‡]
[‡] Georgia Institute of Technology
[†] Samsung AI Center, Toronto
[◇] ILCC, School of Informatics, University of Edinburgh
yjfllpyym@gmail.com federico.f@samsung.com
bonnie@inf.ed.ac.uk dyang888@gatech.edu

Abstract

The availability of corpora has led to significant advances in training semantic parsers in English. Unfortunately, for languages other than English, annotated data is limited and so is the performance of the developed parsers. Recently, pretrained multilingual models have been proven useful for zero-shot cross-lingual transfer in many NLP tasks. What else does it require to apply a parser trained in English to other languages for zero-shot cross-lingual semantic parsing? Will simple language-independent features help? To this end, we experiment with six Discourse Representation Structure (DRS) semantic parsers in English, and generalize them to Italian, German and Dutch, where there are only a small number of manually annotated parses available. Extensive experiments show that despite its simplicity, adding Universal Dependency (UD) relations and Universal POS tags (UPOS) as model-agnostic features achieves surprisingly strong improvement on all parsers. We have publicly released our code at https://github.com/GT-SALT/Multilingual-DRS-Semantic-Parsing.

1 Introduction

Semantic parsing is the task of transducing natural language to meaning representations, which in turn can be expressed through many different semantic formalisms including Discourse Representation Theory (DRT) (Kamp and Reyle, 2013), Abstract Meaning Representation (AMR) (Banarescu et al., 2013), and so on. However, manually annotating meaning representations in a new language is a painstaking process which explains why there are only a few datasets available for different formalisms in languages other than English. For instance, in the case of DRT, even in the latest release of the Parallel Meaning Bank (PMB) (v3.0, Abzianidze et al., 2017), there is no gold training data available for Italian and Dutch and the few annotated sentences are used as development and test data. How can one train a parser when training data is unavailable? This work answers this question by looking at what it is required for zero-shot cross-lingual semantic parsing – learning a semantic parser for English and testing it on other languages.

Prior research on cross-lingual semantic parsing leveraged machine translation techniques to map the semantics from a language to another (Damonte and Cohen, 2018). However, these methods require parallel corpora to extract automatic alignments which are often noisy or not available at all. Other work exploited parameter-shared models, which are based on language-independent representations. In particular, cross-lingual word embeddings and pretrained multilingual models (Hu et al., 2020) have been used in various cross-lingual NLP tasks, including semantic parsing (Oepen et al., 2020; Sherborne et al., 2020). However, pretrained multilingual models are computationally expensive while cross-lingual word embeddings have inferior performance. To this end, we propose to add simple language-independent features in zero-shot cross-lingual semantic parsing, which are *lightweight extensions* to all models instead of designing new architectures with high time and space complexity.

Specifically, we explore cross-lingual syntactic features, Universal Dependencies (UD) and Universal POS tags (UPOS) (De Marneffe et al., 2014), which have been widely annotated in 90 languages. Since semantic parsers need to understand the syntax of sentences well, and UD relations and UPOS have been shown to be strong indicators of semantic roles (Reddy et al., 2017), we hypothesize that using UD and UPOS as simple language-independent and model-agnostic features can boost the performance of zero-shot cross-lingual semantic parsing in all models and languages. To test

†Work done while Federico Fancellu was at the University of Edinburgh.

Gold DRS: DRS(schläfrig(S1) TIME(S1 T1) THEME(S1 Z0) Speaker(Z0) Time(T1) EQ(T1 Y0) Now(Y0))

DRS before adding UD: DRS(ich(X1) bin(X2) TIME(E1 T1) THEME(E1 X1) Time(T1) EQ(T1 Y0) Now(Y0))

DRS after adding UD: DRS(Person(X1) schläfrig(S1) TIME(S1 T1) THEME(S1 X1) Time(T1) EQ(T1 Y0) Now(Y0))

Figure 1: The Discourse Representation Structure (DRS) for "*Ich bin schläfrig. (I am sleepy.)*", including ground truth DRS, parsed DRS before and after adding UD relations as features.

our hypothesis, we focus on the PMB, where sentences in English, German, Italian and Dutch are annotated with their meaning representations. The annotations in the PMB are based on Discourse Representation Theory (DRT, Kamp and Reyle, 2013). Figure 1 shows an annotation example usign DRT for the sentence "*Ich bin schläfrig. (I am sleepy.)*". A Discourse Representation Structure (DRS) is a nested structure with unary and binary predicates representing semantic roles, alongside logic operators (e.g. ¬) and discourse relations (e.g. CONTINUATION). In this example one can see that although a parser could understand the coarse meaning of the whole sentence without UD, it struggles with understanding some lexical-level meaning. However, it successfully identifies "*schläfrig*" as the event, based on its UD relation tag "root".

We carry out our experiments on 6 DRS parsers to test different architectures (LSTM vs. Transformers) as well as different decoding strategies (sequential vs. coarse-to-fine). Whereas the original parsers utilize a sequential neural encoder with monolingual representations, we experiment with cross-lingual representations (i.e. cross-lingual word embeddings and a multilingual pretrained encoder), and language-independent features (i.e. UPOS, UD relations and structures). We also use tree-based encoders to replace the sequential encoder, in order to assess whether modelling syntax is beneficial. Results show that adding UD relations and UPOS as features, despite its frustrating simplicity, leads to surprisingly strong zero-shot cross-lingual semantic parsers, even when UD are the only input used during encoding. Also, UD further boost the performance of strong pretrained multilingual models. Surprisingly, small non-pretrained well-designed coarse-to-fine decoding models with UD relations and UPOS even outperform large pretrained multilingual models in some languages.

A DRS is usually represented in a nested 'box' structure. However, the systems we reference in this paper all use a linearization of this box representation. To avoid any confusion, we only show examples of linearized DRS throughout the paper.

2 Method

2.1 Models

Our models are all encoder-decoder architectures that take as input a natural language sentence $S = s_1...s_{|S|}$ and output a linearized DRS L as a sequence of tokens $y_1...y_{|L|}$. Each model differ in the particular encoder or decoder used, as described in the remainder of this section.

Coarse-to-fine Decoding Models: Our first set of models are based on the coarse-to-fine encoder-decoder architecture of Liu et al. (2018).

Encoder. We first experiment with a BiLSTM encoder (**C2F-BiLSTM**). In cross-lingual settings, we concatenate cross-lingual embeddings with UD relation embeddings and optionally Universal POS embeddings as model input. To model the dependency structure directly, we also test with a child-sum tree-LSTM (Tai et al., 2015) as an alternative to the BiLSTM (**C2F-TreeLSTM**), where each word in the input sentence corresponds to a node in the dependency tree. However, completely discarding word order and context in TreeLSTM might hurt performance. Thus, we combine tree-LSTM and BiLSTM to get **C2F-Bi/TreeLSTM**, where tree-LSTM inputs are initialized using the last layer of a Bi-LSTM (Chen et al., 2017).

Decoder. At decoding time we follow Liu et al. (2018) in reconstructing the linearized DRS representations in three steps coarse-to-fine, each conditioned on the previous one: first, we predict the outer DRS tags which correspond to the semantic environment (e.g. the 'boxes') that predicates will be placed in. Then, the system predicts unary and binary predicates, and in the last step their arguments. The decoder also makes use of a copying mechanism to predict those predicates that are also lemmas in the input sentence (e.g. "*schläfrig*"). We refer to the reader to the original paper for more details.

Sequential Decoding Models: van Noord et al. (2018) introduced another way of losslessly linearizing DRS, along with a dedicated parser. In-

	Model	German			Italian			Dutch		
		P	R	F	P	R	F	P	R	F
PMB 2.1	C2F-BiLSTM-B$_W$	0.608	0.503	0.551	0.619	0.608	0.614	0.555	0.510	0.532
	C2F-BiLSTM$_{W,D}$	0.689	0.651	0.669	0.749	0.727	0.735	0.652	0.620	0.635
	C2F-BiLSTM$_{W,P,D}$	**0.728**	**0.675**	**0.700**	**0.774**	**0.742**	**0.758**	**0.694**	**0.640**	**0.666**
	C2F-BiLSTM$_D$	0.652	0.643	0.647	0.711	0.695	0.703	0.624	0.602	0.613
	C2F-TreeLSTM$_W$	0.513	0.465	0.488	0.581	0.574	0.578	0.551	0.516	0.533
	C2F-TreeLSTM$_{W,D}$	0.702	0.657	0.679	0.749	**0.722**	0.735	0.678	0.597	0.635
	C2F-TreeLSTM$_{W,P,D}$	**0.748**	**0.684**	**0.715**	**0.769**	0.716	**0.742**	**0.701**	**0.615**	**0.655**
	C2F-TreeLSTM$_D$	0.666	0.666	0.666	0.723	0.686	0.704	0.652	0.585	0.617
	C2F-Bi/TreeLSTM$_W$	0.640	0.566	0.601	0.676	0.628	0.651	0.625	0.552	0.586
	C2F-Bi/TreeLSTM$_{W,D}$	0.670	0.664	0.667	0.741	0.689	0.714	0.641	0.583	0.611
	C2F-Bi/TreeLSTM$_{W,P,D}$	**0.726**	**0.684**	**0.704**	**0.763**	**0.730**	**0.746**	**0.675**	**0.616**	**0.644**
	LSTM-N$_W$	0.491	0.514	0.503	0.537	0.583	0.559	0.507	0.519	0.513
	LSTM-N $_{W,D}$	**0.617**	**0.577**	**0.596**	**0.675**	**0.671**	**0.673**	**0.597**	**0.585**	**0.591**
	Transformer-N$_W$	0.527	0.543	0.535	0.559	0.628	0.592	0.535	0.574	0.554
	Transformer-N$_{W,D}$	**0.599**	**0.606**	**0.602**	**0.651**	**0.681**	**0.666**	**0.602**	**0.605**	**0.603**
	XLM-R-Enc-N	0.645	0.651	0.648	0.647	0.673	0.660	0.687	0.690	0.688
	XLM-R-Enc-N$_D$	**0.653**	**0.653**	**0.653**	**0.652**	**0.674**	**0.663**	**0.696**	**0.693**	**0.695**
PMB 3.0	LSTM-N$_W$	0.425	0.457	0.440	0.457	0.507	0.480	0.446	0.495	0.469
	LSTM-N$_{W,D}$	**0.474**	**0.520**	**0.496**	**0.517**	**0.559**	**0.537**	**0.482**	**0.526**	**0.503**
	Transformer-N$_W$	0.452	0.448	0.450	0.478	0.541	0.508	0.481	0.514	0.497
	Transformer-N$_{W,D}$	**0.480**	**0.503**	**0.491**	**0.545**	**0.583**	**0.563**	**0.524**	**0.557**	**0.540**
	XLM-R-Enc-N	0.678	0.697	0.687	0.658	0.677	0.668	0.732	0.730	0.731
	XLM-R-Enc-N$_D$	**0.719**	**0.712**	**0.715**	**0.692**	**0.709**	**0.701**	**0.759**	**0.756**	**0.757**

Table 1: Results of adding language-independent features in various cross-lingual models in PMB 2.1 and 3.0, where *W*, *P* and *D* are the cross-lingual word embeddings, POS embeddings and Dependency relation embeddings respectively. Bold fonts indicate the best performance of features with the same model.

stead of a bracketed representation, predicates, logic operators and discourse relations, along with their arguments, are represented as a sequence of tokens separated by a special symbol "|||". Each variable argument is represented as a relative index pointing to its referent, if already introduced. We refer the reader to the original paper for more details.

To generate such linearized meaning representations, we use either a word-level LSTM encoder-decoder model with copying mechanism (**LSTM-N**) or alternatively, we follow Liu et al. (2019) to replace the LSTM with a Transformer encoder-decoder architecture with copying mechanism (**Transformer-N**). Given the competitive performances of multilingual pretrained models in zero-shot NLP tasks recently (Hu et al., 2020), we also experiment with XLM-R, a state-of-the-art pretrained multilingual model. To adapt such NLU model in our encoder-decoder semantic parser, we use XLM-R to initialize encoder and randomly initialize the Transformer decoder (**XLM-R-Enc-N**).

2.2 Language-Independent Features

In order to make the model directly transferable to the German, Italian and Dutch test data, we use both (1) **UD relations and structure** (*D*) based on UD parses for English, German, Italian and Dutch and (2) **Universal POS tags** (*P*)(Petrov et al., 2011). Both features are extracted using UDPipe (Straka and Straková, 2017).

	TT/min	PT/s	Size/P
XLM-R-Enc-N	300	156	773.3M
C2F-BiLSTM$_D$	9(33x)	25(6.2x)	2.5M(309x)
C2F-BiLSTM$_{W,D}$	10(30x)	27(5.8x)	168.7M(4.6x)

Table 2: Model size and running time in PMB 2.1, where TT is training time till convergence, PT is Parsing/Inference time in German test set and Size is the model size measured with number of Parameters (P).

All the models are based on either of the following cross-lingual representations: (1) **Cross-lingual word embeddings**(*W*) where we use the MUSE (Conneau et al., 2017) pre-trained cross-lingual word embeddings; (2) **Multilingual pre-trained model** where XLM-R (Conneau et al., 2019) is used , considering that XLM-R performs better than XLM (Lample and Conneau, 2019), and mBERT (Wu and Dredze, 2019) in most cross-lingual NLU tasks.

3 Experiments and Results

Data and Evaluation We use the PMB v.2.1.0 for the first series of experiments, where coarse-to-fine decoding models can be used. The dataset consists of 4405 English, 1173 German, 633 Italian and

	operators			non-lexical predicate (*u*)			non-lexical predicate (*b*)			lexical predicate		
	P	R	F	P	R	F	P	R	F	P	R	F
German	0.716	0.378	0.495	0.770	0.664	0.713	0.563	0.528	0.545	0.729	0.733	0.731
Italian	0.930	0.385	0.544	0.797	0.773	0.785	0.563	0.612	0.586	0.662	0.771	0.712
Dutch	0.583	0.189	0.286	0.762	0.665	0.710	0.564	0.543	0.553	0.483	0.607	0.538

Table 3: Error analysis of cross-lingual C2F-BiLSTM in PMB 2.1, where *u* and *b* represent unary and binary.

583 Dutch sentences. We divide the English sentences into 3072 training, 663 development and 670 testing sentences. We consider all the sentences in other languages as test set. We also conduct experiments using PMB v.3.0, where presuppositions and sense tags are considered. However, these additional tags are not compatible with the bracketed structure used in coarse-to-fine decoding models and therefore we use v.3.0 to evaluate sequential decoding models. For English, there are 6620 training, 885 development and 898 testing sentences. There are 403 German, 547 Italian and 483 Dutch sentences in test sets. We use Counter (Van Noord et al., 2018) to evaluate the performance of our models. Counter uses a random hill climbing graph-matching algorithm to look for the best alignment between the predicted and gold DRS and computes precision, recall and F1. For further details, the reader can refer to Van Noord et al. (2018).[1]

3.1 Using Language-Independent Features in Coarse-to-fine Decoding Models

To explore the roles of various language-independent features in zero-shot cross-lingual semantic parsing, we use C2F-BiLSTM as baseline and compare it to C2F-TreeLSTM and C2F-Bi/TreeLSTM. We also conducted ablation studies on the features used. As shown in Table 1, we found that:

(1) **UD relation features are crucial for zero-shot cross-lingual semantic parsing**. Adding UD relations significantly improves the performance in all three coarse-to-fine decoding models in all three languages, compared to using cross-lingual word-embedding alone. Models using UD relation embeddings alone (*D*) perform well, given that the performance does not drop much after deleting cross-lingual word embeddings. However, modeling UD structure via tree encoders does not help zero-shot cross-lingual semantic parsing consistently.

(2) **UPOS features further boost the performance**. After adding UPOS (*P*), all coarse-to-fine models perform even better, reaching state-of-the-art, though the improvement is not as large as adding UD to cross-lingual word embeddings.

3.2 Using UD Relations in All Models

In Table 1, we also examine whether the large improvement made by UD relations is agnostic to models, by adding them in all six baseline models in German, Italian and Dutch. We found that **UD relations lead to consistent and robust improvements.** Results in PMB 2.1 show that adding UD relation features improves the performance in all three languages and six models. Although XLM-R-Enc-N model performs better than non-pretrained models with only cross-lingual embeddings, simple non-pretrained coarse-to-fine models with UD features outperform that large pretrained multilingual model in German and Italian. Considering non-pretrained models require much less model size and training/inference time in Table 2, using UD features in deliberately designed non-pretrained models has its advantage over larger pretrained multilingual models. As for PMB 3.0, UD relation features improve the performance in all models and languages as well, and the improvement is significant even in the pretrained multilingual model. Such improvement is also consistent regardless of different ways of evaluation and linearizing DRS meaning representations in PMB 2.1 and 3.0.

3.3 Error Analysis

We use C2F-BiLSTM, the best cross-lingual model in PMB 2.1, to perform error analysis to assess the quality of the prediction for *operators* (i.e. logic operators like "Not" as well as discourse relations "Contrast"), *unary non-lexical predicates* (e.g. time(t)), *binary non-lexical predicates* (e.g. Agent(e,x)), and *lexical predicates* (e.g. open(e)). Results in Table 3 show that predicting operators and binary predicates is hard, compared to the other two categories. Prediction of lexical predicates is relatively good even though most tokens in the test set were never seen during training. This can be attributed to the copying mechanism that transfers tokens from the input directly during predication.

1. Detailed preprocessing and evaluation are in Appendix.

4 Related work

Previous work have explored two main methods for cross-lingual semantic understanding. One method requires parallel corpora to extract alignments between source and target languages using machine translation (Padó and Lapata, 2005; Damonte and Cohen, 2017; Zhang et al., 2018; Xu et al., 2020), often followed by projection of semantic representations (Reddy et al., 2017). The other method is to use parameter-shared models based on cross-lingual representations such as cross-lingual word embeddings (Duong et al., 2017; Susanto and Lu, 2017; Mulcaire et al., 2018; Hershcovich et al., 2019; Cai and Lapata, 2020), pretrained multilingual models (Zhu et al., 2020; Li et al., 2020; Oepen et al., 2020), and universal POS tags (Blloshmi et al., 2020). Recently, Ozaki et al. (2020); Samuel and Straka (2020); Dou et al. (2020) conducted supervised German DRS parsing with pretrained multilingual models, but they did not explore zero-shot cross-lingual semantic parsing. Besides, although UD was proven useful in other cross-lingual tasks (Subburathinam et al., 2019), it has been under-explored in cross-lingual semantic parsing.

5 Conclusion

This work proposes to use simple language-independent features for the task of zero-shot cross-lingual semantic parsing. We show that simple UD and UPOS features can significantly improve the performance of cross-lingual semantic parsers based on coarse-to-fine decoding techniques or pretrained multilingual models. In the future, we plan to use such features for other semantic formalisms (e.g. AMR) and other languages (e.g. Chinese).

Acknowledgments

We thank the members of Georgia Tech SALT group for their feedback on this work. This work is supported in part by grants from Amazon and Salesforce. We gratefully acknowledge the support of NVIDIA Corporation with the donation of GPU used for this research.

Federico Fancellu contributed to this work prior to joining the Samsung AI Centre Toronto, as a Postdoc at the University of Edinburgh. The views expressed (or the conclusions reached) are their own and do not necessarily represent the view of Samsung Research America, Inc.

References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik Van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The parallel meaning bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. *arXiv preprint arXiv:1702.03964*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Rexhina Blloshmi, Rocco Tripodi, and Roberto Navigli. 2020. Enabling cross-lingual amr parsing with transfer learning techniques. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2487–2500.

Rui Cai and Mirella Lapata. 2020. Alignment-free cross-lingual semantic role labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3883–3894.

Huadong Chen, Shujian Huang, David Chiang, and Jiajun Chen. 2017. Improved neural machine translation with a syntax-aware encoder and decoder. *arXiv preprint arXiv:1707.05436*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Marco Damonte and Shay B Cohen. 2017. Cross-lingual abstract meaning representation parsing. *arXiv preprint arXiv:1704.04539*.

Marco Damonte and Shay B Cohen. 2018. Cross-lingual abstract meaning representation parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1146–1155.

Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–92.

Longxu Dou, Yunlong Feng, Yuqiu Ji, Wanxiang Che, and Ting Liu. 2020. Hit-scir at mrp

2020: Transition-based parser and iterative inference parser. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 65–72.

Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip Cohen, and Mark Johnson. 2017. Multilingual semantic parsing and code-switching. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 379–389.

Daniel Hershcovich, Zohar Aizenbud, Leshem Choshen, Elior Sulem, Ari Rappoport, and Omri Abend. 2019. Semeval-2019 task 1: Cross-lingual semantic parsing with ucca. *arXiv preprint arXiv:1903.02953*.

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 generalization. *arXiv preprint arXiv:2003.11080*.

Hans Kamp and Uwe Reyle. 2013. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*, volume 42. Springer Science & Business Media.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2020. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark. *arXiv preprint arXiv:2008.09335*.

Jiangming Liu, Shay B Cohen, and Mirella Lapata. 2018. Discourse representation structure parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 429–439.

Jiangming Liu, Shay B Cohen, and Mirella Lapata. 2019. Discourse representation structure parsing with recurrent neural networks and the transformer model. In *Proceedings of the IWCS Shared Task on Semantic Parsing*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Phoebe Mulcaire, Swabha Swayamdipta, and Noah Smith. 2018. Polyglot semantic role labeling. *arXiv preprint arXiv:1805.11598*.

Stephan Oepen, Omri Abend, Lasha Abzianidze, Johan Bos, Jan Hajic, Daniel Hershcovich, Bin Li, Tim O'Gorman, Nianwen Xue, and Daniel Zeman. 2020. Mrp 2020: The second shared task on cross-framework and cross-lingual meaning representation parsing. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 1–22.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Hiroaki Ozaki, Gaku Morio, Yuta Koreeda, Terufumi Morishita, and Toshinori Miyoshi. 2020. Hitachi at mrp 2020: Text-to-graph-notation transducer. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 40–52.

Sebastian Padó and Mirella Lapata. 2005. Cross-linguistic projection of role-semantic information. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 859–866. Association for Computational Linguistics.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Siva Reddy, Oscar Täckström, Slav Petrov, Mark Steedman, and Mirella Lapata. 2017. Universal semantic parsing. *arXiv preprint arXiv:1702.03196*.

David Samuel and Milan Straka. 2020. \'ufal at mrp 2020: Permutation-invariant semantic parsing in perin. *arXiv preprint arXiv:2011.00758*.

Tom Sherborne, Yumo Xu, and Mirella Lapata. 2020. Bootstrapping a crosslingual semantic parser. *arXiv preprint arXiv:2004.02585*.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Ananya Subburathinam, Di Lu, Heng Ji, Jonathan May, Shih-Fu Chang, Avirup Sil, and Clare Voss. 2019. Cross-lingual structure transfer for relation and event extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 313–325.

Raymond Hendy Susanto and Wei Lu. 2017. Neural architectures for multilingual semantic parsing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 38–44.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

Rik Van Noord, Lasha Abzianidze, Hessel Haagsma, and Johan Bos. 2018. Evaluating scoped meaning representations. *arXiv preprint arXiv:1802.08599*.

Rik van Noord, Lasha Abzianidze, Antonio Toral, and Johan Bos. 2018. Exploring neural methods for parsing discourse representation structures. *Transactions of the Association for Computational Linguistics*, 6:619–633.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual nlu. *arXiv preprint arXiv:2004.14353*.

Sheng Zhang, Kevin Duh, and Benjamin Van Durme. 2018. Cross-lingual semantic parsing. *arXiv preprint arXiv:1804.08037*.

Qile Zhu, Haidar Khan, Saleh Soltan, Stephen Rawls, and Wael Hamza. 2020. Don't parse, insert: Multilingual semantic parsing with insertion based decoding. *arXiv preprint arXiv:2010.03714*.

Model	P	R	F
C2F-BiLSTM$_W$	0.8698	0.8361	0.8526
C2F-TreeLSTM$_W$	0.8342	0.8016	0.8176
C2F-BiLSTM$_{W,D}$	0.8764	**0.8593**	0.8678
C2F-TreeLSTM$_{W,D}$	0.8569	0.8356	0.8461
C2F-BiLSTM$_D$	0.6629	0.6417	0.6521
C2F-TreeLSTM$_D$	0.6550	0.6589	0.6569
C2F-BiLSTM$_{W,P,D}$	**0.8919**	0.8584	**0.8748**
C2F-TreeLSTM$_{W,P,D}$	0.8590	0.8362	0.8474
LSTM-N	0.8317	0.8041	0.8177
Transformer-N	0.8264	0.8126	0.8194
XLM-R-Enc	0.8579	0.8283	0.8428
LSTM-N	0.8018	0.7567	0.7786
Transformer-N	0.8109	0.7615	0.7854
XLM-R-Enc	0.8377	0.7955	0.8161
XLM-R-Enc + *D*	**0.8407**	**0.7988**	**0.8192**

Table 4: Results for monolingual semantic parsing (i.e. trained and tested in English). PMB 2.1 results are above solid line, PMB 3.0 results are under solid line.

A Monolingual DRS Parsing

For completeness, along with the results for the cross-lingual task, we also report results for monolingual English semantic parsing in Table 4. Coarse-to-fine models are still state-of-the-art for monolingual semantic parsing for English in PMB 2.1, where DRS can be converted to tree-based representations. Dependency features in conjunction with word and PoS embeddings lead to the best performance; however, in all settings explored, treeLSTMs do not outperform BiLSTMs. In PMB 3.0, models with pretrained encoders and Dependency features perform best.

B Preprocessing and Evaluation of PMB 2.1 and PMB 3.0

PMB data can be downloaded from `https://pmb.let.rug.nl/data.php`.

Unlike other work on the PMB (e.g. van Noord et al., 2018), Liu et al. (2018) does not deal with presupposition due to constraints in converting DRS meaning representation to tree-based representations. In PMB 2.1, presupposed variables are extracted from a main box and included in a separate one. We revert this process so to ignore presupposed boxes. Similarly, we also do not deal with sense tags. For fair comparison, all other models are using the same preprocessed meaning representation in PMB 2.1.

In PMB 3.0, presuppositions and sense tags are considered. Thus, coarse-to-fine decoding models can not be used.

Note that lexical predicates in PMB are in English, even for non-English languages. Since this is not compatible with copying mechanism in coarse-to-fine decoding models, LSTM-N and Transformer-N, we revert predicates to their original language in PMB 2.1 by substituting them with the lemmas of the tokens they are aligned to. In PMB 3.0, we do not conduct such reversion, which is compatible with XLM-R-Enc-N, where there is no copying mechanism.

C Model Details

Coarse-to-fine models are based on the coarse-to-fine encoder-decoder architecture of Liu et al. (2018). In order to be used as input to the parser, Liu et al. (2018) first convert the DRS into tree-based representations, which are subsequently linearized into PTB-style bracketed sequences. We use the same conversion. For further details about the conversion and the model, we refer the reader to the original paper.

LSTM-N, Transformer-N and XLM-R-Enc is based on van Noord et al. (2018)'s way of losslessly linearizing DRS meaning representations and the corresponding parser. Clauses in DRS are represented as sequences without changing the order, where a special symbol "|||" is used to start a new clause and variables in clauses are represented as relative indices. Based on sentences and such linearized meaning representations, we use word-level encoder-decoder models to generating parses. We refer the reader to the original paper for further details.

D Training Details

In all models, UD relation embeddings (*D*) and POS tag embeddings (*P*) are randomly initialized and updated during training. Cross-lingual word embeddings are fixed during training and XLM-R is updated.

We adapted OpenNMT (Klein et al., 2017) for LSTM-N and Transformer-N models, while used fairseq (Ott et al., 2019) to implement XLM-R-Enc model.

We manually tune the hyper-parameters. For coarse-to-fine LSTM models, we use two layer BiLSTM or TreeLSTM in the encoder side. We use dropout with 0.5 as dropout rate and Adam optimizer with a learning rate of 5e-4. For LSTM-N, we use batch size 12 and the SGD optimizer with an initial learning rate of 0.7. The learning rate is decayed by 0.7 every 1500 steps. For Transformer-N, we use 6-layer Transformer. We also use a batch

size of 512 tokens and the Adam optimizer with an initial learning rate of 1e-3. The learning rate is decayed by 0.9 every 1000 steps.

In the XLM-R-Enc-N model, we use XLM-R base model to initialize the encoder and randomly initialize a 12-layer transformer decoder. We use Adam as optimizer. Inspired by Liu and Lapata (2019), we use a larger learning rate on the decoder side, in order to solve the discrepancy between pretrained encoder and non-pretrained decoder. Specifically, in the encoder side, we use a polynomial learning rate scheduler with 2e-5 as max learning rate and 5000 warmup steps. In the decoder side, we use a polynomial learning rate scheduler with 5e-5 as max learning rate and 2500 warmup steps. Total update steps are all 50000. We use a label smoothing rate 0.1 and dropout rate 0.1.

We train all models on GPU GeForce RTX 2080. The training time for Coarse-to-fine models, Transformer-N and LSTM-N models are all within 2 hours. The training time for XLM-R-Enc-N is 5-6 hours. The number of parameters in Coarse-to-fine models, Transformer-N and LSTM-N models are similar to the parameters of C2F-BiLSTM shown in Table 2 in the paper. The number of parameters in XLM-R-Enc-N is shown in Table 2 in the paper. We use English development data for evaluation, where the validation performance is similar to the performance in the English test set reported in Table 1.