Towards a Unified Multi-Domain Multilingual Named Entity Recognition Model

Mayank Kulkarni[*#2], **Daniel Preoţiuc-Pietro**[1], **Karthik Radhakrishnan**[1], **Genta Indra Winata**[1], **Shijie Wu**[1], **Lingjue Xie**[1], **Shaohua Yang**[*]
[1]Bloomberg [2]Amazon Alexa AI
{dpreotiucpie, gwinata}@bloomberg.net

Abstract

Named Entity Recognition is a key Natural Language Processing task whose performance is sensitive to choice of genre and language. A unified NER model across multiple genres and languages is more practical and efficient through leveraging commonalities across genres or languages. In this paper, we propose a novel setup for NER which includes multi-domain and multilingual training and evaluation across 13 domains and 4 languages. We explore a range of approaches to building a unified model using domain and language adaptation techniques. Our experiments highlight multiple nuances to consider while building a unified model, including that naive data pooling fails to obtain good performance, that domain-specific adaptations are more important than language-specific ones and that including domain-specific adaptations in a unified model can reach performance close to training multiple dedicated monolingual models at a fraction of their parameter count.

1 Introduction

Identifying named entities, such as organization and people in text is a key NLP task situated upstream of other NLP tasks such as co-reference resolution (Ratinov and Roth, 2012; Dutta and Weikum, 2015; Miwa and Bansal, 2016; Luo and Glass, 2018) or relation extraction (Nguyen and Grishman, 2015; Zhong and Chen, 2021) and can enhance applications including information retrieval (Carpineto and Romano, 2012; Berger and Lafferty, 2017) and summarization (Cheng and Lapata, 2016; Liu and Lapata, 2019; Maddela et al., 2022; Hofmann-Coyle et al., 2022). However, it has been established that NER is very sensitive to genre differences (Augenstein et al., 2017; Agarwal et al., 2021),[1] with models trained on one genre performing poorly on a different one. As a result, multiple data sets were created to allow domain-specific models to be trained. Yet, domain adaptation especially across multiple genres has shown the promise that a single model could improve performance on multiple domains (Wang et al., 2020; Liu et al., 2021). Further, the advent of pretrained multilingual models (Wu and Dredze, 2019; Conneau et al., 2020) enables transfer across languages in an straightforward way by simply feeding heterogeneous language data in fine-tuning, making cross-lingual training feasible and a new dimension of adaptation available for exploration. Thus, performance on a specific genre and language pair could be improved by leveraging commonalities in training data across both genre and language dimensions, which is enabled by the significant amount of annotated data sets that are publicly available.

The main research question becomes *what is the best way to leverage data from different languages and genres for NER* in this multilingual multi-domain setup? In this paper, we attempt to answer this question by using data sets available across multiple genres and languages to improve

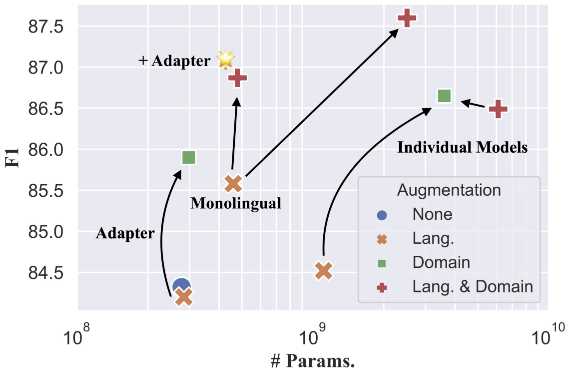

Figure 1: A graphic comparison of performance of language, genre and joint adaptation by demo parameter count.

[*] Work done while at Bloomberg.

[#]The authors are listed in alphabetical order.

[1]Throughout this paper, by genre we refer to a collection of documents with variations in style or structure that might impact modeling (Santini et al., 2006); we use “domain" interchangeably with “genre" when referring to modeling concepts.

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2210–2219
May 2-6, 2023

performance across all data sets. To this end, we compile a collection of 22 data sets across 4 different languages and spanning 13 domains. To the best of our knowledge, this is the first attempt to building a unified multi-domain multilingual NER model.

We empirically explore our core research question through several experiments. First, we aim to identify whether sharing models or parts of the model across languages, domains or both is more beneficial in training. In general, simply pooling all the available data is likely sub-optimal as domain-specific differences in named entity mentions are useful to model, although using more data is usually more beneficial and can lead to improved robustness of the model. We explore several sharing techniques on top of state-of-the-art transformer-based encoders such as data pooling and mixture of experts methods, previously effective in cross-lingual learning, and language or domain specific adapter heads. Our results (Fig. 1) show that sharing genre information across languages is much more beneficial for performance than sharing language information across genres for all types of adaptation techniques.

Next, we compare monolingual encoders like RoBERTa, which can provide a better representation for text in each language, and multilingual encoders like XLM-R, which enables knowledge sharing from multiple languages, as starting points for fine-tuning NER models when genre and language annotated data is available. We find that the monolingual models pooling all the data from a particular language perform best and outperform their cross-lingual counterparts.

Throughout, we explore the trade-offs between the total number of model parameters and performance, which can bring practical benefits in terms of reduced maintenance and increased efficiency. We find that doing domain adaptation using adapter heads achieves a good trade-off between performance and parameter count and could represent the optimal solution in deploying a unified model.

Our contributions are as follows:

- Introducing the multilingual multi-domain NER setup;
- Extensive experiments on **13 domains** and **4 languages** using a variety of models and adaptation methods which highlight the best unified model architecture and show that modeling domains is more effective than languages;
- Analysis of the performance / efficiency trade-offs.

2 Data Sets

We create a collection of 22 data sets across 4 languages and 13 unique genres. For English, we use CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003), Filings (Salinas Alvarado et al., 2015), OntoNotes-En (Hovy et al., 2006) with 6 genres (Pradhan et al., 2013; Wang et al., 2020) and Twitter (Ritter et al., 2011); for Chinese, we use MSRA (Levow, 2006), OntoNotes-Zh (Hovy et al., 2006) with 6 genres and Weibo (Peng and Dredze, 2015, 2016; He and Sun, 2017); for German, we use CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003), Legal (Leitner et al., 2019) and Wikipedia (Balasuriya et al., 2009); for Spanish we use CoNLL-2002 (Tjong Kim Sang, 2002) and Wikipedia (Balasuriya et al., 2009). Note that not all languages contain the same genres and not all genres are present in each language, although there is overlap between genres and languages.

2.1 Statistics

We have a total of 502,720 training examples with 109,657 for validation and 105,255 for testing. We consider the following entity types: Person (`PER`), Organization (`ORG`), Location (`LOC`), and Miscellaneous (`MISC`), either removing extra types or collapsing them into the overarching parent entity class. We maintain the train/dev/test splits for all of these data sets and evaluate on test. Tab. 1 shows the number of training, validation and testing data points across each of the languages when domains in the language are combined. Here we can see that despite German and Spanish having few domains, the number of data points are in fact more than English and Chinese which have more domains.

Language	Train	Dev	Test
English	103,121	22,344	22,557
Chinese	67,129	15,672	11,314
German	234,297	50,471	50,611
Spanish	98,173	21,170	20,773
Total	502,720	109,657	105,255

Table 1: Data set statistics per language.

2.2 Entity-Type Mapping

The datasets do not have identical entity types. Thus, we apply pre-processing to standardize the

Language	Domain	ORG	PER	LOC	Dropped
English	Ritter OntoNotes	company N/A	person, musicartist N/A	geo-loc N/A	N/A TIME, CARDINAL, NORP, DATE, ORDINAL, QUANTITY, MONEY, PRODUCT, EVENT, PERCENT, WORK_OF_ART, LAW, LANGUAGE
German	Legal	UN, INN, GRT, MRK	RR, AN	LD, ST, STR, LDS	GS, RS, VO, LIT, VS, VT, EUN
Chinese	Weibo OntoNotes	ORG.NAM, ORG.NOM N/A	PER.NAM, PER.NOM PERSON	LOC.NAM, LOC.NOM GPE	N/A EVENT, NORP, TIME, FAC, QUANTITY, MONEY, CARDINAL, ORDINAL, LOC, LAW, WORK_OF_ART, PERCENT, LANGUAGE, PRODUCT

Table 2: Entity-type Mapping across data sets.

labels. Tab. 2 illustrates the pre-processing to map entities to the ORG, PER, LOC and MISC types. We do not list the simple mapping when the ORG, PER, LOC types exist themselves and blank spaces signify no mapping was done for the type. We reference previous work to map types to our subsets and also refer to the original data set paper to infer type mappings. Additionally, if a type does not map to any of the 4 types we train and evaluate on we drop the type as seen in the last column of the table.

3 Methods

To answer our core research question, we explore several methods inspired by approaches from both multilingual NER (Al-Rfou et al., 2015; Rahimi et al., 2019; Tedeschi et al., 2021) and multi-domain NER (Liu et al., 2020; Wang et al., 2020).

3.1 Multilingual Encoders

Pretrained multilingual encoders learn strong multilingual representation. In particular, we use XLM-R base (Conneau et al., 2020), a strong multilingual encoder.

3.1.1 Individual Models

The models consist of XLM-R base as the encoder, followed by the sequence tagging head in the form of a linear layer.

- **Data Pooling.** We fine-tune 1 model by naively pooling data from all languages and domains;
- **Per Lang.** We fine-tune 4 models using all data from each of the 4 languages;
- **Per Dom.** We fine-tune 1 model per domain using all data across all languages for that domain, resulting in 13 models (e.g., one CoNLL model trained using English, German and Spanish);
- **Per Lang. & Dom.** We fine-tune one model for each language and domain resulting in 22 models (e.g., CoNLL English, CoNLL German).

3.1.2 Mixture of Expert Models

Past multilingual NER research showed promising results using Mixture of Expert (MoE) (Shazeer et al., 2017) based models. MoE models are built on the premise that a set of experts can be parametrically learnt based on the training data without any explicit notion of matching an expert to a specific language or domain. MoE based models could be trained with a regular training setup (Jacobs et al., 1991), with gradient reversal methods (Ganin and Lempitsky, 2015) or with an adversarial loss (Chen and Cardie, 2018; Chen et al., 2019). We train MoE with regular training setup.

Given encoder output H for a sequence of length M, we introduce N experts, $E_i = FFN_i(H_t)$ with one hidden layer, where $i \in \{1, \dots, N\}, t \in \{1, \dots, M\}$, and a linear domain/language N-class classifier $C_{D/L} = Softmax(W_{C_{D/L}} H_{CLS})$. We take $\sum_i \alpha_i E_i$ with α_i from $C_{D/L}$ and feed it to a shared NER classifier. Thus experts are assigned at the sequence-level.[2] These are jointly trained in a multi-task learning setup with a cross-entropy NER loss $\mathcal{L}_{\text{NER}}^{\sum_i \alpha_i E_i}$ associated with all experts E_i and a Domain **(Dom. MoE)** or Language **(Lang. MoE)** prediction cross-entropy loss $\mathcal{L}_{D/L}$, yielding $\mathcal{L}_{\text{NER}}^{\sum_i \alpha_i E_i} + \mathcal{L}_{D/L}$. The loss is backpropagated through all the experts, both NER and domain/language classifier and the shared encoder.

3.1.3 Adapter Models

Research in multi-domain NER has found that adding private layers that are updated by data from

[2]We also experimented with token-level expert assignment, but observe worse results on the dev set.

each domain and shared layers updated by data from all domains is an effective way to improve multi-domain performance (Wang et al., 2020). Similar to the private layers explored in multi-domain NER, Adapters (Pfeiffer et al., 2020; Lin et al., 2020; Winata et al., 2021) used in conjunction with Transformer-based models demonstrated promise in further boosting the performance. We thus introduce adapter heads on top of the encoder. We leave variants of adapters that lie within each layer of the encoder (Houlsby et al., 2019; He et al., 2022) as future work.

The adapters A_i use the same model architecture (as MoE models), but are only updated by data from a given domain or language. It is equivalent to MoE with a predefined expert assignment. Fig. 2 shows the architecture of the adapter models used in our experiments.

Thus, for a given data point D_i the loss is computed as $\mathcal{L}_{\mathrm{NER}}^{A_i}$ and only backpropagated through A_i, the NER classifier and the shared encoder.

- **Lang. Adp.** We create 4 adapter heads, one for each language and use the gold language label to pick the adapter;
- **Dom. Adp.** We create 13 adapter heads, one for each domain and use the gold domain label to pick the adapter;
- **Dom. Adp + DP.** We create 13 adapter heads and employ an auxiliary Domain Prediction objective $\mathcal{L}_{D/L}$ during training;
- **Dom. Adp + DP + SA.** In addition, we add a shared head which is updated for all examples, similar to the shared/private setup in (Wang et al., 2020) for multi-domain adaptation.

While adapters in each layer with frozen encoder performs on par with fine-tuning all parameters (Houlsby et al., 2019), it does not outperform it eithers. Thus, we also update the transformer layers as part of the training process. We also explored combining language and domain adapters but this resulted in worse performance and we omit it for brevity.

3.2 Monolingual Encoders

Finally, we explore monolingual encoders, which can provide a better representation of each language but are not able to transfer knowledge across languages. We identify monolingual BERT/RoBERTa versions for each of the 4 languages: English (Liu et al., 2019), Chinese (Cui et al., 2020), Spanish (de la Rosa et al., 2022), and German[3].

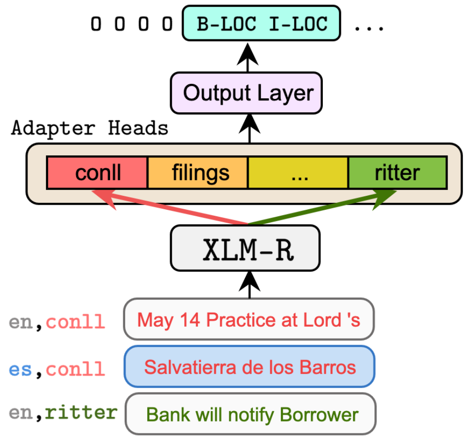

Figure 2: Our NER tagger with XLM-R encoder and domain adapters. Texts and adapter heads are color-coded to indicate the heads used for each domain.

- **Per Lang.** We fine-tune each monolingual encoder with all the NER data from the corresponding language, resulting in 4 models;
- **Per Lang. & Dom.** We fine-tune one model for each domain based on monolingual encoder, resulting in 22 models;
- **Dom. Adp.** We add domain adapters (**Dom. Adp.**, §3.1.3) to monolingual encoder. This results in 4 models, one for each language, with the number of adapters equal to the number of domains in each language.

3.3 Hyperparameters

We use the open-source Transformers library (Wolf et al., 2020) to facilitate reproducibility. For all experiments, we use a learning rate of 1e-5 on the AdamW optimizer (Loshchilov and Hutter, 2017), with no warm up, a batch size of 32 trained across 50 epochs on an NVIDIA V100 GPU. We use the same hyperparameters across all experiments to allow for comparability.

4 Results

We evaluate the models listed in §3 on all data sets. Tab. 3 illustrates the results averaged across languages obtained using F1 calculated with the CoNLL evaluation script. Granular results for each individual genre and language are in App. B. We treat the data pooling method in multi-lingual encoders as our baseline in terms of performance and number of parameters. Our findings are as follows:

[3]The model is taken from https://www.deepset.ai/german-bert

Models	# param	en	zh	de	es	Avg.
# Domains		9	8	3	2	
`Initiate with XLM-R (multilingual, base)`						
Data Pooling	×1.00	82.12	75.70	89.51	89.96	84.32
Per Lang.	×4.00	+0.54	+0.67	-0.27	-0.15	+0.2
Per Dom.	×13.00	<u>+3.99</u>	+4.45	+0.47	<u>+0.40</u>	+2.33
Per Lang. & Dom.	×22.00	+3.53	+4.27	<u>+0.49</u>	+0.38	+2.17
Lang. MoE	×1.02	+0.12	+0.47	0	+0.07	+0.17
Dom. MoE	×1.07	+0.51	+1.03	-0.36	-0.33	+0.21
Lang. Adp	×1.02	-0.09	+0.33	-0.37	-0.35	-0.12
Dom. Adp	×1.07	+1.95	+4.62	+0.26	-0.53	+1.58
+ DP	×1.07	+1.65	+3.49	-0.15	-0.83	+1.04
+ DP + SA	×1.08	+1.60	+3.66	-0.30	-0.01	+1.24
`Initiate with Monolingual RoBERTa (base)`						
Per Lang.	×1.65	+1.94	+3.16	-0.24	+0.16	+1.26
Per Lang. & Dom.	×9.03	**+4.23**	<u>+6.55</u>	**+1.28**	**+1.05**	**+3.28**
Dom. Adp	×1.73	+3.05	**+6.61**	+0.22	+0.30	<u>+2.55</u>

Table 3: Results in macro-F1 for each language averaged across all domains within the language and overall average across the four languages. Number of parameters are relative to Data Pooling. **Bold** and <u>underline</u> indicate the best and second best performing models.

Domain vs. Language: In Fig. 1, we observe that across all types of methods (individual models, MoE and adapters), training models that leverage information about domain across languages is more beneficial when compared to sharing information across different genres in the same language, with gains of up to 1.70 – 2.13 F1. We hypothesize this result is due to the well documented sensitivity of NER to nuances specific to genres (Augenstein et al., 2017) such as entity distribution, document structure or capitalization patterns, whereas multilingual models manage to better preserve this information across languages. In addition, domain-specific models even perform slightly better than language- and domain-specific models (+0.16).

Adapters vs. MoE: When comparing methods, we observe that MoE techniques provide limited gains over data pooling (0.17–0.21) contrary to past cross-lingual experiments. The adapter heads provide bigger improvement compared to MoE with same number of parameters, while using shared layers and domain predictors as in multi-domain adaptation (Wang et al., 2020) fails to further boost performance. However, both adaptation strategies lag behind training domain specific models (+0.75), which however come with a much larger number of parameters (up to ×20) and added maintenance cost when deployed.

Monolingual vs. Multilingual Models: The monolingual results demonstrate that, if available, these models lead to better performance than their multilingual counterparts (+1.06 and +1.09 when comparing similar setups), which is natural as they have a better representation of each language. We find that the domain adapter method offers a good trade-off between performance (-0.73) and model size (×0.18), outperforming all models that perform adaptation across languages.

Impact of domain diversity: Finally we also observe that English and Chinese have much more diversity because of the number of domains, thus adding more capacity through the domain adaptation results in improved performance. However, since German and Spanish have fewer domains but an equal if not more training data points, we find that adding capacity is not necessarily helpful.

5 Conclusion and Future Work

This paper introduces the first extensive evaluation of multilingual multi-domain NER using a collection of 22 data sets spanning 4 languages. Through a series of experiments, we demonstrate that genre information is more important to be shared, even across languages, than sharing information from other genres in the same language. However, these cross-lingual methods are outperformed by domain adaptation over genres in monolingual models, if these models are available. We also explored trade-offs between model size and performance, showing that adapter heads strike a good balance, offering relatively little reduction in performance for an order of magnitude less parameters. For future work, we will explore additional experimental setups that include testing on domains or languages where limited or no data is available for training.

Limitations

Our research focuses on high-resource languages where annotated NER data sets and pretrained language models are available from only two language families. We have yet to explore how these findings translate to low resource languages or languages where annotated data sets are not available. We note that there are more domains available for English and Chinese, and since we are computing macro-F1 scores, the results over-emphasize performance on these languages, although Spanish and German show similar result patterns. Additionally, we only use 4 entity types (i.e., PER, ORG, LOC, MISC) across all datasets by dropping the other entities. Finally, due to limited computing resources and large number of experiments, we experiment with XLM-R base and thus do not compare with state-of-the-art results for each of these individual language/domain results which are usually obtained using XLM-R large.

Ethics Statement

We use publicly available data sets in our experiments with permissive licenses for research experiments. We do not release new data or annotations as part of this work.

Acknowledgements

We are grateful to Xinyu Hua for feedback on a draft of this manuscript. We sincerely thank the three anonymous reviewers for their insightful comments on our paper.

References

Oshin Agarwal, Yinfei Yang, Byron C. Wallace, and Ani Nenkova. 2021. Interpretability analysis for named entity recognition to understand system predictions and how they can improve. *Computational Linguistics*, 47(1):117–140.

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, ICDM, pages 586–594.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R. Curran. 2009. Named entity recognition in Wikipedia. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web)*, pages 10–18, Suntec, Singapore. Association for Computational Linguistics.

Adam Berger and John Lafferty. 2017. Information retrieval as statistical translation. 51(2):219–226.

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, 44(1):1–50.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112, Florence, Italy. Association for Computational Linguistics.

Xilun Chen and Claire Cardie. 2018. Multinomial adversarial networks for multi-domain text classification. 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 1226–1240, New Orleans, Louisiana. Association for Computational Linguistics.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Javier de la Rosa, Eduardo G Ponferrada, Paulo Villegas, Pablo González de Prado Salas, Manu Romero, and Marıa Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68:13–23.

Sourav Dutta and Gerhard Weikum. 2015. Cross-document co-reference resolution using sample-based clustering with knowledge enrichment. *Transactions of the Association for Computational Linguistics*, 3:15–28.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, ICML, pages 1180–1189.

Hangfeng He and Xu Sun. 2017. F-score driven max margin neural network for named entity recognition in Chinese social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 713–718, Valencia, Spain. Association for Computational Linguistics.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*, ICLR.

Ella Hofmann-Coyle, Mayank Kulkarni, Lingjue Xie, Mounica Maddela, and Daniel Preoţiuc-Pietro. 2022. Extractive entity-centric summarization as sentence selection using bi-encoders. 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)*, AACL-IJCNLP, pages 326–333.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, ICML, pages 2790–2799.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained named entity recognition in legal documents. In *International Conference on Semantic Systems*, pages 272–287.

Gina-Anne Levow. 2006. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, Sydney, Australia. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2020. Exploring versatile generative language model via parameter-efficient transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 441–459, Online. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2020. Zero-resource cross-domain named entity recognition. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 1–6, Online. Association for Computational Linguistics.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35 of *AAAI*, pages 13452–13460.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*, ICLR.

Hongyin Luo and Jim Glass. 2018. Learning word representations with cross-sentence dependency for end-to-end co-reference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4829–4833, Brussels, Belgium. Association for Computational Linguistics.

Mounica Maddela, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2022. EntSUM: A data set for entity-centric extractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3355–3366, Dublin, Ireland. Association for Computational Linguistics.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48, Denver, Colorado. Association for Computational Linguistics.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for Chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554, Lisbon, Portugal. Association for Computational Linguistics.

Nanyun Peng and Mark Dredze. 2016. Improving named entity recognition for Chinese social media with word segmentation representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 149–155, Berlin, Germany. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2012. Learning-based multi-sieve co-reference resolution with knowledge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1234–1244, Jeju Island, Korea. Association for Computational Linguistics.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90, Parramatta, Australia.

Marina Santini, Richard Power, and Roger Evans. 2006. Implementing a characterization of genre for automatic genre identification of web pages. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 699–706, Sydney, Australia. Association for Computational Linguistics.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, ICLR.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Jing Wang, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2020. Multi-domain named entity recognition with genre-aware and agnostic inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8476–8488, Online. Association for Computational Linguistics.

Genta Indra Winata, Guangsen Wang, Caiming Xiong, and Steven Hoi. 2021. Adapt-and-adjust: Overcoming the long-tail problem of multilingual speech recognition. In *22nd Annual Conference of the International Speech Communication Association*, Interspeech, page 361.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

Appendix

A Data Set Details

In this section, we provide details about the statistics of the data sets, our hypothesis on what makes them challenging tasks and also pre-processing we perform to allow for reproducible results.

A.1 Statistics

A more in-depth look at the distributions of the domains across languages can be seen in Tab. 4 for German, Tab. 5 for Spanish, Tab. 8 for English, and Tab. 9 for Chinese. The tables show that English has the most diverse set of domain distribution, followed by Chinese, with a bulk of the data coming from MSRA, German, where Legal and Wiki constitute a large amount and Spanish, which is largely dominated by Wiki. The more diverse set of domains makes the language more challenging to achieve a consistently high average score, which is also evident in our results.

	conll2003	legal	wiki
Train (%)	5.19	19.93	74.88
Dev (%)	5.68	19.83	74.49
Test (%)	5.94	19.78	74.28

Table 4: Domain Distribution for German data sets.

	conll	wiki
Train (%)	8.48	91.52
Dev (%)	9.05	90.95
Test (%)	7.31	92.69

Table 5: Domain Distribution for Spanish data sets.

B Fine-grained Results

In Tab. 3, we see the averaged results across domains for each language, however it is not easy to infer the performance on each for any given language. In an effort to provide more transparency, we provide the performance for each domain within a given language in Tab. 6, Tab. 7, Tab. 10, and Tab. 11.

Models/Domain	conll2003	legal	wiki
`Initialized with XLM-R Multilingual`			
Data Pooling	81.24	96.05	91.23
Per Lang.	80.61	95.80	91.30
Per Dom.	82.08	96.41	91.45
Per Lang. and Dom.	82.26	96.41	91.33
Lang. MoE	81.28	96.10	91.14
Dom. MoE	80.42	95.94	91.09
Lang. Adp	80.33	95.92	91.17
Domain Adp	81.87	96.18	91.25
+ DP	81.41	95.81	90.87
+ DP + SA	81.18	95.68	90.77
`Initialized with Monolingual RoBERTa`			
Per Lang.	81.00	95.55	91.25
Per Lang. & Dom.	84.69	96.05	91.61
Dom. Adp	82.08	95.77	91.36

Table 6: Fine-grained results for domains within German.

Models/Domain	conll2002	wiki
`Initialized with XLM-R Multilingual`		
Data Pooling	86.72	93.20
Per Lang.	86.51	93.10
Per Dom.	87.46	93.26
Per Lang. and Dom.	87.59	93.09
Lang. MoE	87.02	93.04
Dom. MoE	86.42	92.84
Lang. Adp	86.27	92.96
Dom. Adp	85.93	92.93
+ DP	85.72	92.55
+ DP + SA	87.08	92.81
`Initialized with Monolingual RoBERTa`		
Per Lang.	86.97	93.26
Per Lang. & Dom.	88.78	93.25
Dom. Adp	87.36	93.15

Table 7: Fine-grained results for domains within Spanish. DP is Domain Prediction and SA indicates shared adapter.

	conll2003	filings	onto_bc	onto_bn	onto_mz	onto_nw	onto_tc	onto_wb	ritter
Train (%)	13.62	1.00	11.00	9.05	5.66	29.32	10.79	14.65	4.92
Dev (%)	14.55	0.99	10.88	8.96	5.59	29.00	10.67	14.49	4.86
Test (%)	15.31	0.98	10.79	8.88	5.55	28.73	10.58	14.36	4.83

Table 8: Domain Distribution for English data sets.

	msra	onto_bc	onto_bn	onto_mz	onto_nw	onto_tc	onto_wb	weibo
Train (%)	53.63	10.51	8.92	4.54	3.86	9.01	7.56	1.97
Dev (%)	57.43	9.65	8.19	4.17	3.54	8.27	6.94	1.81
Test (%)	40.94	13.37	11.36	5.78	4.92	11.47	9.63	2.52

Table 9: Domain Distribution for Chinese data sets.

Models/Domain	conll2003	filings	bc	bn	mz	nw	tc	wb	ritter
`Initialized with Multilingual XLM-R`									
Data Pooling	87.21	88.59	88.11	90.92	89.10	91.60	70.69	70.87	61.94
Per Lang.	87.84	86.10	89.14	91.51	89.13	92.13	71.78	71.83	64.45
Per Dom.	92.00	95.48	89.85	92.48	90.81	93.14	73.29	76.44	71.46
Per Lang. & Dom.	91.25	95.48	89.35	92.53	91.19	92.79	73.49	75.85	68.96
Lang. MoE	87.20	88.37	88.44	90.91	88.51	91.74	71.30	70.59	63.13
Dom. MoE	87.59	87.50	87.99	91.35	88.23	91.80	72.18	72.42	64.62
Lang. Adp	87.02	88.33	87.72	91.29	88.26	91.77	70.37	70.54	63.01
Dom. Adp	90.36	90.98	89.19	92.38	89.66	92.22	71.59	75.23	65.02
+ DP	90.27	89.45	88.67	92.34	89.55	91.93	71.21	76.44	64.07
+ DP + SA	90.17	89.46	88.56	92.24	89.18	91.96	72.80	75.50	63.60
`Initialized with Monolingual RoBERTa`									
Per Lang.	88.81	90.66	89.80	91.92	89.61	92.48	72.29	73.76	67.22
Per Lang. & Dom.	91.69	94.44	89.81	92.85	91.15	93.42	73.77	75.78	74.25
Dom. Adp	91.85	89.71	89.64	93.07	90.79	92.97	72.29	75.94	70.30

Table 10: Fine-grained results for domains within English. DP is Domain Prediction and SA indicates shared adapter.

Models/Domain	msra	bc	bn	mz	nw	tc	wb	weibo
`Initialized with XLM-R Multilingual`								
Data Pooling	81.09	78.30	79.08	72.75	84.19	81.18	68.63	60.35
Per Lang.	80.85	76.58	78.94	74.30	84.68	83.33	68.36	63.90
Per Dom.	91.81	77.62	82.42	76.25	90.23	84.59	73.51	64.74
Per Lang. & Dom.	91.82	77.82	82.21	76.07	89.74	83.81	73.59	64.74
Lang. MoE	81.04	77.05	79.40	72.47	84.91	84.59	68.35	61.56
Dom. MoE	79.59	76.76	79.32	74.21	85.66	85.29	70.21	62.81
Lang. Adp	80.57	79.07	78.51	73.46	84.62	82.99	68.10	60.89
Dom. Adp	89.45	79.47	82.14	76.21	89.99	84.86	74.34	66.10
+ DP	89.36	77.08	80.67	75.17	89.76	85.22	74.00	62.24
+ DP + SA	89.00	77.73	81.14	75.82	89.54	85.55	72.16	63.97
`Initialized with Monolingual RoBERTa`								
Per Lang.	82.25	79.78	81.02	75.43	86.39	86.10	71.85	68.04
Per Lang. & Dom.	93.55	80.36	84.05	78.52	91.35	85.84	74.73	69.6
Dom. Adp	93.17	80.34	83.88	78.79	91.22	86.74	76.34	67.96

Table 11: Fine-grained results for domains within Chinese. DP is Domain Prediction and SA indicates shared adapter.