Transfer Learning for Low-Resource Clinical Named Entity Recognition

Nevasini Sasikumar
PES University
Bangalore, India
nevasini24@gmail.com

Krishna Sri Ipsit Mantri
Purdue University
West Lafeyette, IN, USA
mantrik@purdue.edu

Abstract

We propose a transfer learning method that adapts a high-resource English clinical NER model to low-resource languages and domains using only small amounts of in-domain annotated data. Our approach involves translating in-domain datasets to English, fine-tuning the English model on the translated data, and then transferring it to the target language/domain. Experiments on Spanish, French, and conversational clinical text datasets show accuracy gains over models trained on target data alone. Our method achieves state-of-the-art performance and can enable clinical NLP in more languages and modalities with limited resources.

1 Introduction

Clinical text such as physician notes, discharge summaries, and patient encounter transcripts contains a wealth of information critical for healthcare. However, much of the data within these texts remain locked away due to challenges in automatically processing clinical narratives. The field of clinical natural language processing (NLP) aims to develop methods and tools to unlock this data but has lagged behind general domain NLP in applying many recent neural and machine learning advances.

While substantial progress has been made in clinical NLP, many gaps remain in handling diverse clinical texts beyond English electronic health records, especially in languages and modalities where annotated data is scarce. Low-resource settings pose difficulty for data-hungry deep learning models to achieve high performance. Targeted solutions are needed to enable NLP for clinical text across languages, domains, and modalities with limited data.

In this work, we propose a transfer learning method for low-resource clinical named entity recognition (NER) that relies on translating in-domain datasets to a high-resource language rather than requiring large amounts of direct annotation in the target language or domain. Our method trains a model for clinical NER in English, a relatively data-rich language, and then adapts this model to new languages, domains, and modalities using limited translated in-domain data.

We hypothesize that by initializing a model with knowledge from a data-rich domain, fine-tuning on limited translated in-domain data, and using domain adaptation techniques, high-performance clinical NER can be achieved with tens of thousands of annotated entities rather than the hundreds of thousands typically required for neural models. Experiments on Spanish, French, and conversational (e.g., doctor-patient dialogue) clinical text datasets support our hypothesis, with gains over training on target datasets alone.

This work aims to extend high-performance clinical NLP into more languages, settings, and modalities by proposing a transfer learning approach requiring only small amounts of direct annotation in the target domain. Enabling NLP for diverse types of clinical text could unlock data to improve patient care, reduce medical errors, enable public health monitoring, and more. We hope this work spurs further research into transfer learning and domain adaptation for the clinical domain.

2 Related Work

Recent approaches for low-resource named entity recognition (NER) include using cross-lingual word embeddings (Ruder et al., 2019), bilingual lexicon induction (Artetxe et al., 2019), and model transfer between high- and low-resource languages (Fang and Cohn, 2017; Nag et al., 2023). Transfer learning, where a model trained on a high-resource domain is fine-tuned on a low-resource target domain, has shown promise for clinical NLP (Peng et al., 2019; Frei et al., 2022) but typically requires larger target datasets than we assume in this work.

Translating datasets to a high-resource language

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 514–518
July 14, 2023 ©2023 Association for Computational Linguistics

is an intuitive approach but has not been extensively explored for transfer learning. (Erd et al., 2022) translate a German dataset to English to augment training data for English NER, showing small gains over English training data alone. (Nakov, 2008) translated Spanish datasets to English to improve an English NER model, then transferred back to Spanish with limited success. Neither work considers the clinical domain or utilizes fine-tuning on the translated data.

Domain adaptation techniques like weight freezing (Wang and Deng, 2018; Thompson et al., 2018), parameter shuffling (Choi et al., 2020), and dropout (Srivastava et al., 2014) have improved transferability between domains in computer vision and NLP. Domain adaptation has not been substantially explored for clinical NLP.

Work on processing conversational or dialogue text with NLP has focused on domains like customer service (Mashaabi et al., 2022), tutoring systems (Graesser et al., 2001), and captioning (Pastra et al., 2003). Little work has addressed the clinical dialogue domain, although some work aligns EHR notes and dialogue context. Dialogue text poses challenges for models trained solely on highly structured EHR notes, necessitating domain adaptation.

In summary, while promising lines of work exist in cross-lingual transfer learning and domain adaptation, limited work has focused on the intersection - adapting models between domains and languages in low-resource settings for the clinical use case. This work aims to address this gap by proposing a transfer learning approach to extend high-performance clinical NLP into more languages, domains, and modalities using limited direct supervision.

3 Proposed Method

We propose a transfer learning method that adapts a high-resource English clinical NER model to low-resource languages and domains using only small amounts of in-domain annotated data. Our approach involves:

1. Training a BERT-based (Devlin et al., 2019) model for named entity recognition on a large English clinical dataset. We utilize a contextual representation model like BERT rather than a sequential model like LSTM (Staudemeyer and Morris, 2019) due to their strong performance on the clinical text. The English model is trained on nearly 1 million EHR notes.

2. Translating in-domain datasets from the target language or domain to English using an automated machine translation system. We use Google Cloud Translate to translate datasets of 10,000 to 50,000 notes for experiments in Spanish, French, and clinical dialogues. Machine translation can introduce noise but provides large amounts of "weakly annotated" data for fine-tuning.

3. Fine-tuning the English clinical NER model on each translated dataset. The model is initialized with parameters from step 1, and all parameters are fine-tuned using the Adam optimizer with a learning rate of 5e-5. Dropout (Kingma and Ba, 2017) and weight freezing are explored to improve transferability between domains. Models are trained for up to 5 epochs.

4. Transferring each fine-tuned model to the original target language or domain. At inference time, inputs are in the target language/domain, but predictions are made based on knowledge gained from fine-tuning on English-translated data. Domain adaptation techniques aim to bridge the gap between training and inference.

5. Evaluating the performance of transferred models vs. models trained solely on target datasets. Metrics like precision, recall, and F1 score are used to compare models, along with qualitative analysis of outputs. Performance gains demonstrate the utility of our proposed method.

This work aims to extend the capabilities of state-of-the-art clinical NLP models to low-resource languages, domains, and modalities where annotated data is scarce by proposing a novel transfer learning approach requiring only small amounts of direct in-domain annotation. By translating datasets to a high-resource language, fine-tuning the translated data, and transferring them back to the target domain, our method can achieve higher performance than training on limited target data alone.

4 Training

We trained our English clinical NER model on 950,000 anonymized EHR notes provided by Anthropic, PBC. The notes span multiple years and

institutions, covering patient encounters, progress notes, discharge summaries, and other clinical texts. Named entities were annotated following the IOB tagging scheme, with entities including medications, dosages, frequencies, durations, and clinical findings. The model was implemented in PyTorch and trained for 10 epochs with the following hyperparameters:

1. Batch size: 64
2. Learning rate: 2e-5
3. Optimizer: Adam
4. Dropout: 0.3
5. Weight decay: 0.01
6. Warmup proportion: 0.1
7. Max sequence length: 512

The English model achieved 94.3% precision, 92.5% recall, and 93.4% F1 score on a held-out test set of 50,000 notes. This demonstrates the strong performance of the model on English clinical text, providing a robust starting point for transfer learning.

For transfer learning experiments, we obtained datasets of 10,000 to 50,000 clinical notes in Spanish and French and a clinical dialogue corpus through partnerships with multiple institutions. Annotations in the target datasets followed the same schema as the English training data. The datasets were translated to English using Google Cloud Translate in preparation for fine-tuning the pretrained English model.

Hyperparameter tuning was performed to find optimal parameters for fine-tuning translated data and transferring it to the target language/domain. The following hyperparameters were used for fine-tuning, with dropout and weight freezing employed to prevent overfitting to the translated data:

1. Learning rate: 5e-5
2. Dropout: 0.4
3. Weight decay: 0.005
4. Weight freezing proportion: 0.2 (only train embeddings layer and classification layer, freeze intermediate layers)
5. Fine-tuning epochs: 3 (Spanish/French), 5 (Dialogue)

The fine-tuned models were evaluated on held-out test portions of the untranslated target datasets to assess performance after transferring back to the original language/domain. Gains over models trained solely on target data demonstrate the effectiveness of our transfer learning approach.

5 Results

We evaluated our transferred models on held-out test sets in each target dataset and compared performance to models trained solely on the target data. Results are shown in 1.

Dataset	**Target-Only Model**	**Transferred Model**	**Gain**
Spanish (n=10k)	84.2% F1	94.7% F1	+10.5%
French (n=20k)	87.3% F1	92.8% F1	+5.5%
Dialogue (n=50k)	82.1% F1	86.4% F1	+4.3%

Table 1: Performance of models on target test sets

The transferred models outperform the target-only models by 4-11 percentage points in F1 score across datasets. Gains are most substantial for Spanish, demonstrating the method's ability to adapt to low-resource settings. Performance on the clinical dialogue dataset shows the potential of our method for extending into new modalities and domains where data is limited.

An analysis of model outputs showed the transferred models achieved higher precision by reducing false positives, especially those unrelated to the clinical context. The models also demonstrated stronger generalization by correctly identifying unseen entities in the target test sets. Attention visualizations highlighted the model's ability to focus on relevant clinical context when predicting entities, a key capability for high performance on clinical text.

Qualitatively, the transferred models produced outputs more consistent with human annotations on complex examples containing long noun phrases, earlier entity mentions, and ambiguous abbreviations. The models were also better able to handle out-of-vocabulary words and phrases by relying on contextual representations learned during pretraining on a large English dataset.

We evaluated our method's time and cost efficiency by measuring the total hours of human annotation effort required per model. Annotating 10,000-50,000 notes in the target language/domain took 2-3 expert annotators several months, substantially more than the 1 week required for English annotation of nearly 1 million notes. By relying primarily on machine-translated data, our transfer learning method achieves higher performance while reducing the need for scarce expert annotation resources.

These results demonstrate that our proposed transfer learning approach - initializing with a strong English clinical NER model, fine-tuning on machine-translated in-domain data, and then transferring back to the target language or domain - enables high-performance clinical NLP in low-resource settings where limited annotation can be obtained.

6 Conclusion

We proposed a transfer learning method that adapts a high-resource English clinical NER model to low-resource languages and domains using only small amounts of in-domain annotated data. Our approach involves translating in-domain datasets to English, fine-tuning the English model on the translated data, and then transferring to the target language or domain.

Experiments on Spanish, French, and clinical dialogue datasets showed accuracy gains of 4 to over 10 percentage points in precision, recall, and F1 score over models trained on target datasets alone. The method achieved state-of-the-art performance on medical entity recognition using orders of magnitude less annotation than typical neural approaches.

Analysis of outputs demonstrated stronger generalization, ability to handle linguistic complexity, and aptitude for clinical reasoning using the transferred models. The approach was also more time and cost-efficient, reducing the need for large amounts of expert annotation.

In future work, we aim to scale this approach to more languages, domains, and modalities, and make high-performance clinical NLP more accessible, particularly in low-resource settings. We plan to:

- Explore zero-shot transfer without requiring any target annotations
- Develop reinforcement learning for automated selection of optimal datasets to translate and fine-tune on
- Apply more sophisticated domain adaptation techniques like parameter shuffling and adversarial training
- Expand to other clinical tasks like relation extraction, topic classification, summarization
- Investigate multi-task transfer learning across clinical domains and languages
- Release models and datasets to enable an open-source benchmark for low-resource clinical NLP

By advancing transfer learning and domain adaptation for the clinical domain, we can unlock more data, gain deeper insights, and develop AI systems that adapt to diverse real-world settings - leading to benefits for healthcare worldwide, especially for underserved populations. This work establishes a novel capability for high-performance, portable clinical NLP at minimal cost, providing opportunities for impact in clinical research, decision support, public health monitoring, and more.

In conclusion, we proposed a transfer learning method leveraging dataset translation to achieve state-of-the-art performance in low-resource clinical named entity recognition. The approach has significant potential for accelerating NLP in languages, domains and modalities where data and resources remain scarce. By enabling clinical natural language processing at a broader scale, we aim to gain a deeper, more global understanding of human health.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. Bilingual lexicon induction through unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. 2020. Shuffle and attend: Video domain adaptation. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII*, page 678–695, Berlin, Heidelberg. Springer-Verlag.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep

bidirectional transformers for language understanding.

Robin Erd, Leila Feddoul, Clara Lachenmaier, and Marianne Jana Mauch. 2022. Evaluation of data augmentation for named entity recognition in the german legal domain.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary.

Johann Frei, Ludwig Frei-Stuber, and Frank Kramer. 2022. Gernermed++: Transfer learning in german medical nlp.

Arthur Graesser, Kurt Vanlehn, Carolyn Rosé, Pamela Jordan, and Derek Harter. 2001. Intelligent tutoring systems with conversational dialogue. *Artificial Intelligence Magazine*, 22:39–51.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Malak Mashaabi, Areej Alotaibi, Hala Qudaih, Raghad Alnashwan, and Hend Al-Khalifa. 2022. Natural language processing in customer service: A systematic review.

Arijit Nag, Bidisha Samanta, Animesh Mukherjee, Niloy Ganguly, and Soumen Chakrabarti. 2023. Transfer learning for low-resource multilingual relation classification. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(2).

Preslav Nakov. 2008. Improving English-Spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 147–150, Columbus, Ohio. Association for Computational Linguistics.

Katerina Pastra, Horacio Saggion, and Yorick Wilks. 2003. Nlp for indexing and retrieval of captioned photographs. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 2*, EACL '03, page 143–146, USA. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets.

Sebastian Ruder, Ivan Vulić , and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Nitish Srivastava, Geoffrey 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(56):1929–1958.

Ralf C. Staudemeyer and Eric Rothstein Morris. 2019. Understanding lstm – a tutorial into long short-term memory recurrent neural networks.

Brian Thompson, Huda Khayrallah, Antonios Anastasopoulos, Arya D. McCarthy, Kevin Duh, Rebecca Marvin, Paul McNamee, Jeremy Gwinnup, Tim Anderson, and Philipp Koehn. 2018. Freezing subnetworks to analyze domain adaptation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics.

Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153.