Cross-Lingual Knowledge Transfer for Clinical Phenotyping

Jens-Michalis Papaioannou*, Paul Grundmann*, Betty van Aken*, Athanasios Samaras† Ilias Kyparissidis‡, George Giannakoulas†, Felix Gers*, Alexander Löser*

*DATEXIS, Berliner Hochschule für Technik (BHT), Germany,

†First Department of Cardiology, AHEPA University Hospital, Aristotle University of Thessaloniki, Greece

‡Laboratory of Medical Physics and Digital Innovation, Aristotle University of Thessaloniki, Greece

contact: michalis.papaioannou@bht-berlin.de

Berliner Hochschule für Technik

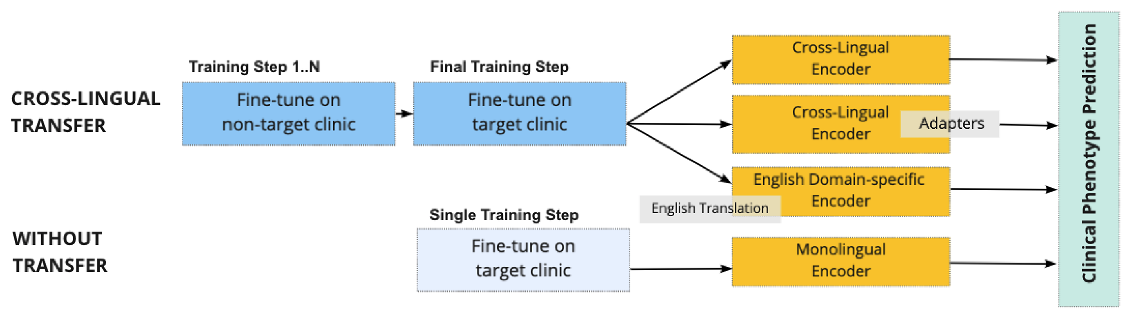
Objectives

1. Automatically categorise patients by clinical phenotype (clinical conditions found in clinical notes)
2. Effective and efficient multilingual data augmentation for low-resource languages without sharing data
3. Use pre-defined CCSR categorization where each category represents a set of ICD codes. Each category may represent a disease or a set of e.g., different arrhythmias or ill defined diseases.

Methods

We restrict our approaches to sequential transfer learning, since it allows to share models across clinics without having to share patient data explicitly.
We compare:

- Cross-Lingual encoder [1, 2] (original language) **XLM-R + Adapter**
- Domain-specific encoder [3] (english translation) **PubMedBERT**, **Spanish B. RoBERTa**
- Monolingual encoders [4, 5, 6] **Spanish BERT**, **GreekBERT**
- Cross-lingual data augmentation (original language)



Datasets

Mimic III - English Language

Mimic III [7] contains de-identified Electronic Health Records (EHR) data including clinical notes in English from the Intensive Care Unit (ICU) of Beth Israel Deaconess Medical Center in Massachusetts between 2001 and 2012.

CodiEsp - Spanish Language

The CodiEsp dataset [8] consists of 1,000 clinical case studies manually selected by doctors and cover a diverse set of medical specialties. The notes are provided in both the original Spanish language and an English translation.

AHEPAcardio - Greek Language

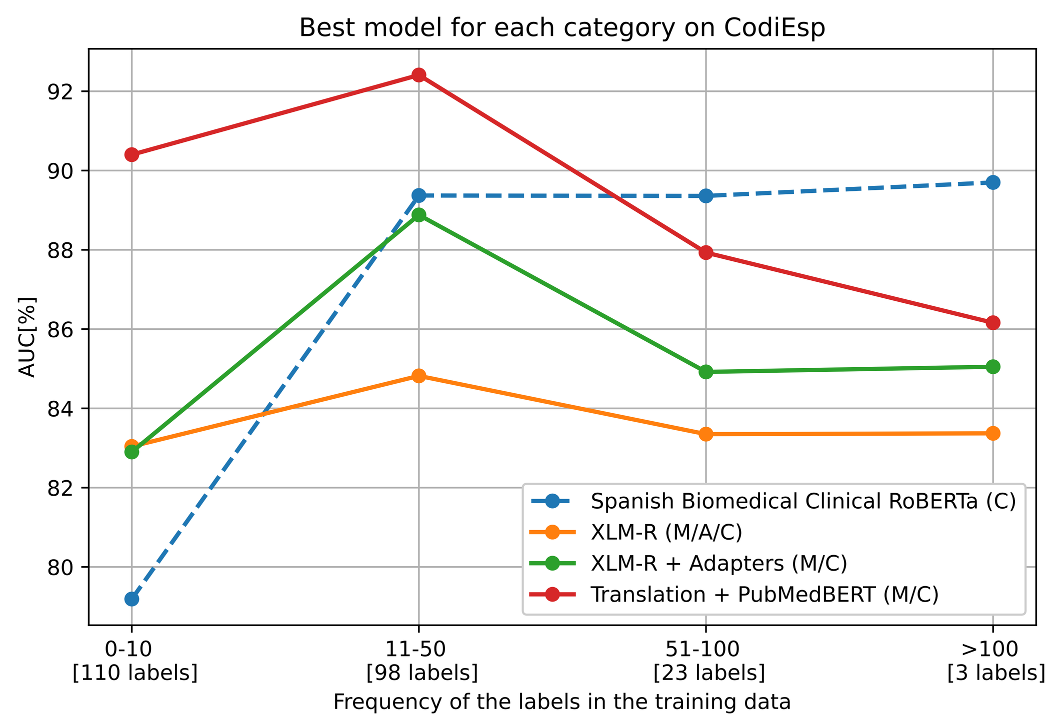
is a collection of around 2,400 discharge summaries and originates from the cardiology clinic of the AHEPA University Hospital in Greece.

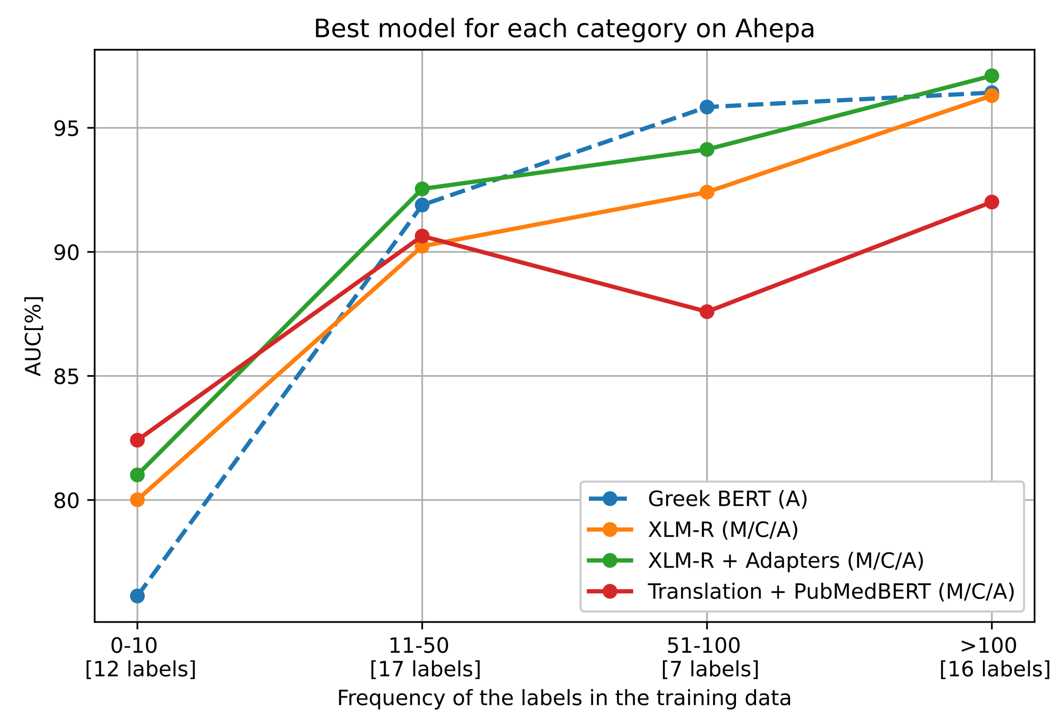
Clinical Note Statistics				
	Train	Dev	Test	Ø Length
CodiEsp	656	165	175	351
Ahepa	1,592	402	393	257
Mimic	24,758	6,187	6,182	649

Acknowledgements

We would like to thank Janine Schleicher and Julius Laufer for their support. Our work is funded by the German Federal Ministry for Economic Affairs and Energy (BMWi) under grant agreement 01MD19003B (PLASS) and 01MK2008MD (Servicemeister).

Main Findings





- Low resource datasets benefit from cross-lingual transfer
- Rare phenotypes gain most out of cross-lingual transfer
- Adapters and translation are both suitable methods for cross-lingual knowledge transfer.
- Adding more data does not necessarily improve results
- Translation quality and translation consistency are important; Abbreviations and style of writing have an impact on translation
- Use adapters when computational complexity is a limiting factor
- If an in-domain translation system is available, translate the text to English and then use an in-domain monolingual encoder

Results

Model	Clinical Phenotyping Macro-AUC [%]	Clinical Phenotyping Macro PR-AUC [%]
Single Dataset Training		
Monolingual Spanish BERT (C)	82.00	25.91
Spanish Biomedical Clinical RoBERTa (C)	84.58	29.89
XLM-R (C)	56.64	5.28
XLM-R + Adapters (C)	61.96	6.43
Translation + PubMedBERT (C_T)	83.45	29.54
Multi Dataset Training		
XLM-R (M → C)	83.52	25.96
XLM-R (M → A → C)	83.82	25.96
XLM-R + Adapters (M → C)	85.63	34.41
XLM-R + Adapters (M → A → C)	83.90	32.22
Translation + PubMedBERT (M → C_T)	**90.95**	**43.13**
Translation + PubMedBERT (M → A_T → C_T)	90.40	41.98

Table 1: Performance for **CodiEsp**. *M*: Mimic, *A*: Ahepa and *C*: CodiEsp. The order represents the fine-tune order. The subscript $_T$ means that the English translation of the texts is used and otherwise the original language. The approach which yields the strongest results is the sequential fine-tuning of the **Domain specific Encoder** first with Mimic and then with the **English translation** of CodiEsp.

Model	Clinical Phenotyping Macro-AUC [%]	Clinical Phenotyping Macro PR-AUC [%]
Single Dataset Training		
Monolingual Greek BERT (A)	90.18	56.22
XLM-R (A)	60.45	12.31
XLM-R + Adapters (A)	56.60	10.30
Translation + PubMedBERT (A_T)	83.15	37.10
Multi Dataset Training		
XLM-R (M → A)	89.87	50.23
XLM-R (M → C → A)	90.03	51.15
XLM-R + Adapters (M → A)	90.15	54.45
XLM-R + Adapters (M → C → A)	**91.50**	**57.63**
Translation + PubMedBERT (M → A_T)	86.20	45.14
Translation + PubMedBERT (M → C_T → A_T)	88.75	49.90

Table 2: Performance for **Ahepa**. *M*: Mimic, *A*: Ahepa and *C*: CodiEsp. The order represents the fine-tune order. The subscript $_T$ means that the English translation of the texts is used and otherwise the original language. The approach which yields the strongest results is the sequential fine-tuning of the **Cross-lingual Encoder plus Adapter** on Mimic, CodiEsp and Ahepa in **original language**.

References

[1] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics, 2020.

[2] Jonas Pfeiffer, Ivan Vulic, Iryna Gurevych, and Sebastian Ruder. MAD-X: an adapter-based framework for multi-task cross-lingual transfer. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7654–7673. Association for Computational Linguistics, 2020.

[3] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *CoRR*, abs/2007.15779, 2020.

[4] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*, 2020.

[5] John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. GREEK-BERT: the greeks visiting sesame street. *CoRR*, abs/2008.12014, 2020.

[6] Casimiro Pio Carrino, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, and Marta Villegas. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario. *CoRR*, abs/2109.03570, 2021.

[7] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad M. Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 2016.

[8] Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. Overview of automatic clinical coding: Annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF ehealth 2020. In Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol, editors, *Working Notes of CLEF 2020 - Conference and Labs off the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020.