School of Computing, Engineering, and the Built Environment
Edinburgh Napier University

PHD STUDENT PROJECT

Application instructions:
Detailed instructions are available at :
https://www.napier.ac.uk/research-and-innovation/doctoral-college/how-to-apply

Prospective candidates are encouraged to contact the Director of Studies (see details below) to discuss the project and their suitability for it.

Project details

Supervisory Team:

- DIRECTOR OF STUDY: Dr Peter Barclay (Email: p.barclay@napier.ac.uk)
- 2ND SUPERVISOR: Prof. Alistair Lawson

Subject Group: Computer Science

Research Areas: Computer Science - Machine Learning

Project Title: Language Models for a Low-resource language: Scottish Gaelic

Project description:

Recent advances in the development of Large Language Models (LLM) have depended on high-levels of computational power and the use of extremely large data sets for training the models. This is a problem when building models for lesser-used languages. For example, Gàidhlig (Scottish Gaelic) is an important cultural resource and heritage language in the UK, but as the current speech community and web presence of the language is comparatively small, there is insufficient data for training language models to a high standard.

This problem can be addressed in a number of ways, including creating new corpora, investigating methods of augmenting the existing training data, or employing transfer learning from models built for related languages (such as the Irish gaBERT).

This project aims to survey the state of the art in Gàidhlig language models, then investigate methods to enhance these models given limited training data, and thereby build and evaluate improved language models, which may be used for

tasks such as named entity recognition, part-of-speech tagging, and dependency parsing.

References:

Barry, J. (2022). Investigating multilingual approaches for parsing universal dependencies [PhD Thesis, Dublin City University]. https://doras.dcu.ie/27698/

Barry, J., Wagner, J., Cassidy, L., Cowap, A., Lynn, T., Walsh, A., Meachair, M. J. Ó., & Foster, J. (2022). gaBERT -- an Irish Language Model (arXiv:2107.12930). arXiv. http://arxiv.org/abs/2107.12930

Batchelor, C. (2019). Universal dependencies for Scottish Gaelic: Syntax. Proceedings of the Celtic Language Technology Workshop, 7–15. https://aclanthology.org/W19-6902.pdf

Chiche, Alebachew, and Betselot Yitagesu. "Part of Speech Tagging: A Systematic Review of Deep Learning and Machine Learning Approaches." Journal of Big Data, vol. 9, no. 1, Jan. 2022, p. 10. BioMed Central, https://doi.org/10.1186/s40537-022-00561-y.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Evans, L., Lamb, W., Sinclair, M., & Alex, B. (2022). Developing Automatic Speech Recognition for Scottish Gaelic. LREC 2022 Workshop Language Resources and Evaluation Conference 20-25 June 2022, 110.

Ghani, Rayid, Rosie Jones, and Dunja Mladenic. "Building Minority Language Corpora by Learning to Generate Web Search Queries." Knowledge and Information Systems, vol.7, no. 1, Jan. 2005, pp. 56–83. DOI.org (Crossref), https://doi.org/10.1007/s10115-003-0121-x

Huang, J., Alex, B., Bauer, M., Salvador-Jasin, D., Liang, Y., Thomas, R., & Lamb, W. (n.d.). A Transformer-Based Orthographic Standardiser for Scottish Gaelic. Retrieved 25 October 2023, from https://sigul-2023.ilc.cnr.it/wp-content/uploads/2023/08/7_Paper.pdf

Maxwell, Mike, and Baden Hughes. "Frontiers in Linguistic Annotation for Lower-Density Languages." Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006, edited by Timothy Baldwin et al., Association for Computational Linguistics, 2006, pp. 29–37. ACLWeb, https://aclanthology.org/W06-0605.

Nguyen, S., & Anderson, C. J. (2023). Do All Minority Languages Look the Same to GPT-3? Linguistic (Mis) information in a Large Language Model. Proceedings of the Society for Computation in Linguistics, 6(1), 400–402.

Streiter, Oliver, et al. "Implementing NLP Projects for Noncentral Languages: Instructions for Funding Bodies, Strategies for Developers." Machine Translation, vol. 20, no. 4, Mar. 2006, pp. 267–89. DOI.org (Crossref),https://doi.org/10.1007/s10590-007-9026-x .

Zhou, Jeffrey, and Neha Verma. Transfer Learning for Low-Resource Part-of-Speech Tagging. 2020.

Candidate characteristics

Education:

Minimum 2:1 degree - Artificial Intelligence, Computer Science, Statistics, Linguistics

Subject knowledge:

Natural Language Processing, Machine Learning

Essential attributes:

- Self-motivated
- Numerate
- Programming Experience

Desirable attributes:

- Working knowledge of Scottish Gaelic