University of Southampton

Human in the Loop AI

Turing at Southampton Research Showcase Event

Dr Stuart E. Middleton
University of Southampton, Electronics and Computer Science, sem03@soton.ac.uk

21st Sept 2022

University of Southampton

Overview

- Speaker
- What is Human in the loop AI?
- Human-in-the-loop NLP
- Summary

Speaker

- Dr Stuart E. Middleton
- Associate Professor, University of Southampton
- Turing Fellow
- Research interests
  - Natural Language Processing (NLP)
  - Information Extraction (IE)
  - Human-in-the-loop NLP
  - Problem areas typically in low training resource and multi-disciplinary environments

University of Southampton

What is Human in the loop AI?

- No Human-in-the-Loop
  - No human in the loop >> AI models trained from datasets
- Human-In-The-Loop AI (HILT) [Middleton 2022]
  - **Human in the training loop** (offline)
    - Humans help AI models to (re)train so they improve over time
  - **Human in the deployment loop** (online)
    - Humans help AI models to perform a task
- Not covered today (but related)
  - **Lifelong learning**
    - Iterative retraining of models over time as it encounters new problems/experiences
  - **Human-machine teaming**
    - Human and AI actors interacting to perform a task (not covered today)

University of Southampton

What is Human in the loop AI?

- Why put a human-in-the-loop anyway when training AI models?
  - Bigger datasets are not always the answer for getting better AI models
  - Researchers often overfit datasets with challenge specific models [Nie 2020]
  - Humans can help AI models move beyond the limitations of static training data
- This talk focusses on **Human-in-the-loop NLP** examples
  - NLP is a sub-field of AI
  - Human-in-the-loop NLP examples today apply to many other sub-fields of AI
  - Machine Learning, Image Processing, Audio Processing, Robotics …

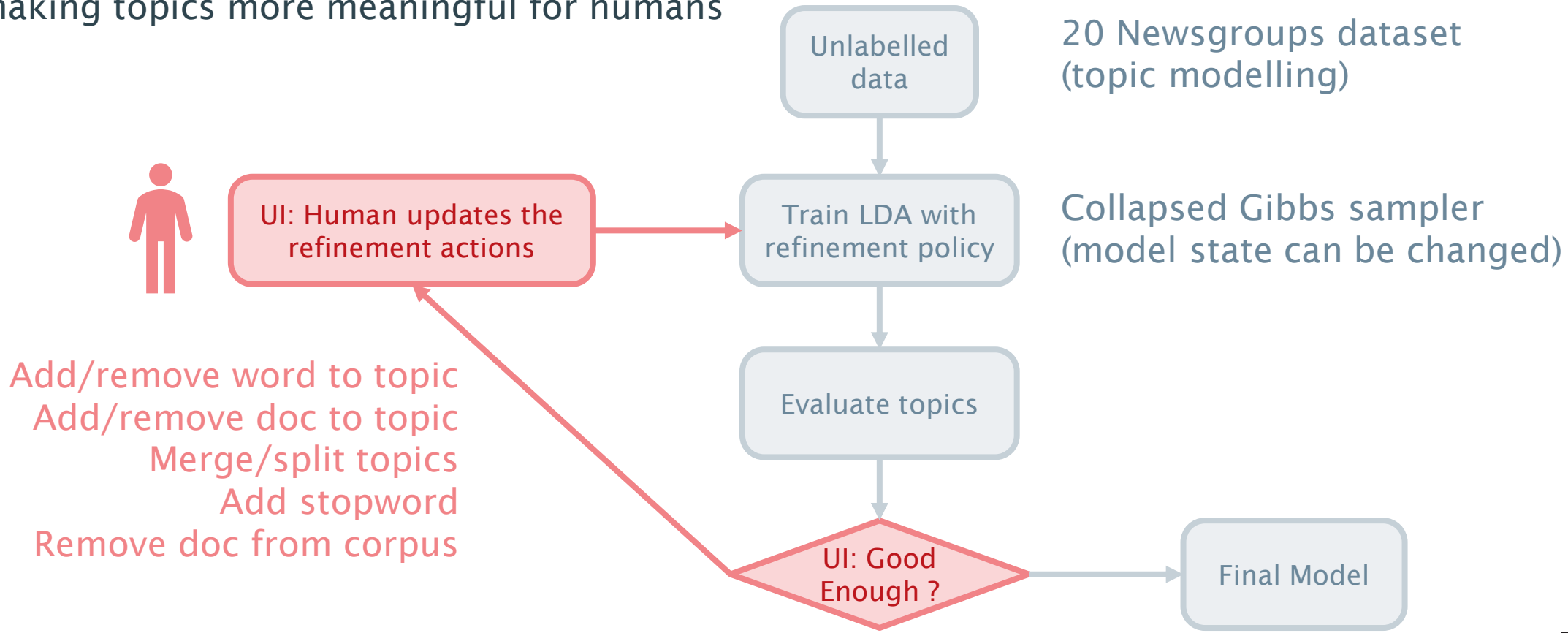
University of Southampton

Human-in-the-loop NLP

- **Interactive sense making** [Bunch 2020] [Middleton 2020]
  - Visualize results >> human analysis >> retrain >> repeat (cycle)
- **Explainable AI** [Hanafi 2020]
  - Explanation model >> human analysis >> retrain >> repeat (cycle)
- **Adversarial training** with human in the loop [Ratner 2017] [Nie 2020]
  - Human defines labelling functions >> train GAN >> GAN infers labels >> train model
  - Human acts like a GAN creates difficult examples >> retrain >> repeat (cycle)
- **Active learning** [Kanchinadam 2020] [Ghai 2020]
  - Model selects unlabelled data for human annotation >> retrain >> repeat (cycle)
- **Few Shot Learning** and **Meta-learning** [Gao 2019] [Yin 2020]
  - Human provides a query and a support set of example classes

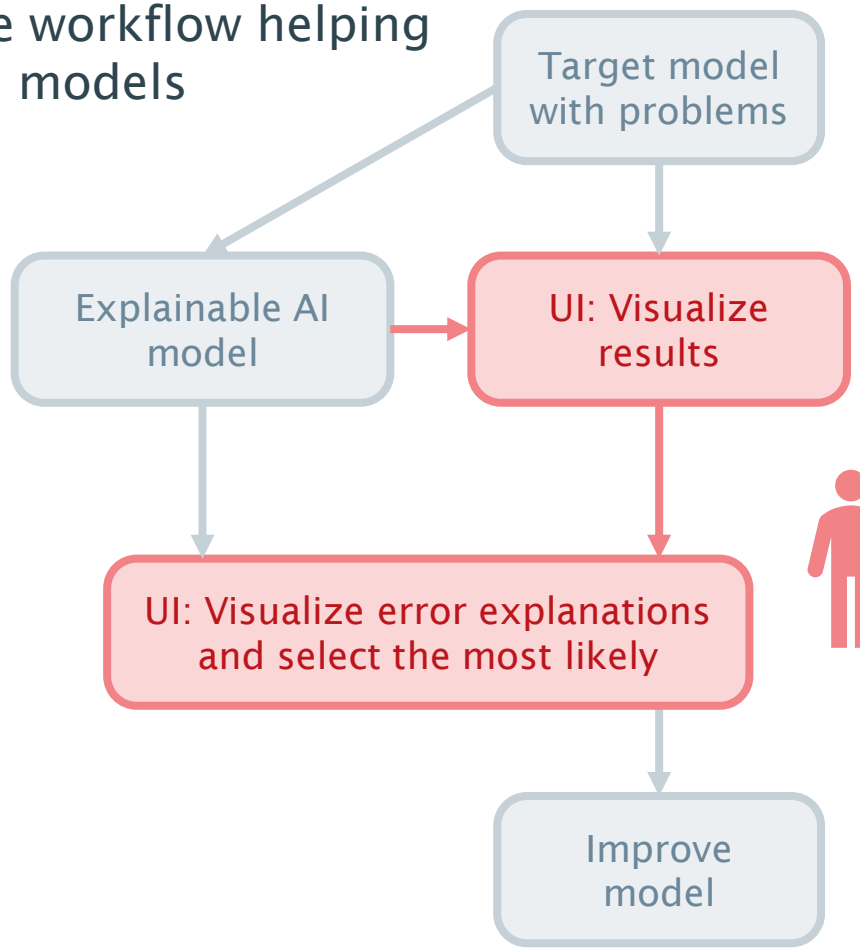
Human-in-the-loop NLP

- Interactive sense making [Bunch 2020]
  - Human-in-the-loop LDA topic modelling making topics more meaningful for humans



Human-in-the-loop NLP

- Explainable AI [Hanafi 2020]
  - Explainable AI interactive workflow helping humans correct errors in models



Data workflow model for a commercial retail system

Error explanations are generated and ranked in order of likelihood based on dataflow coverage etc.

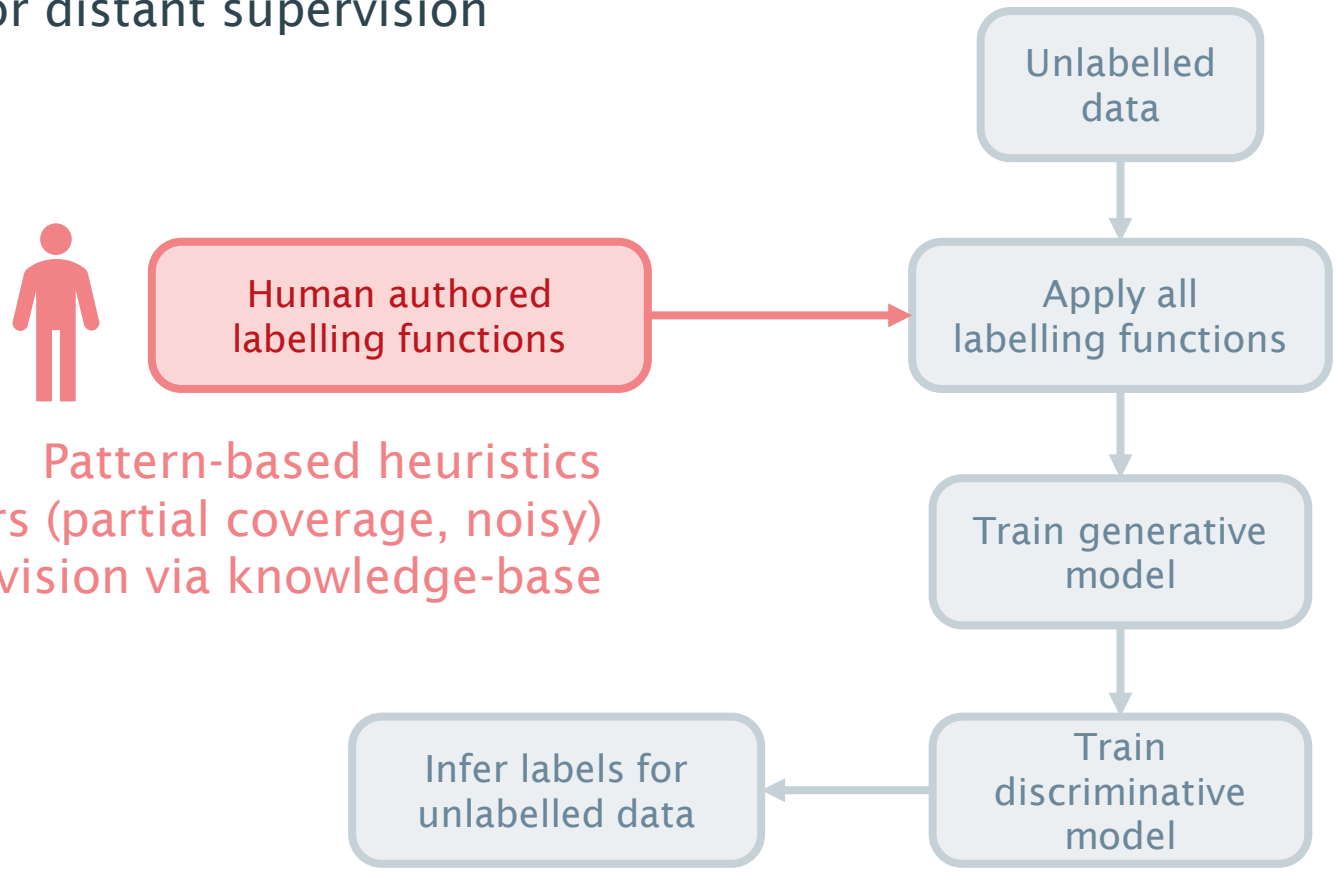
Random sample of dataflows shown to users to identify errors

Users pick most likely error explanation (feedback to fix target model)

Domain-independent explainable AI methods include SHAP, LIME etc.

Human-in-the-loop NLP

- Adversarial training with human-in-the-loop [Ratner 2017]
  - Human authored labelling functions used to train a GAN for distant supervision



Pattern-based heuristics
Weak classifiers (partial coverage, noisy)
Distant supervision via knowledge-base

Pubmed dataset (relation extraction)

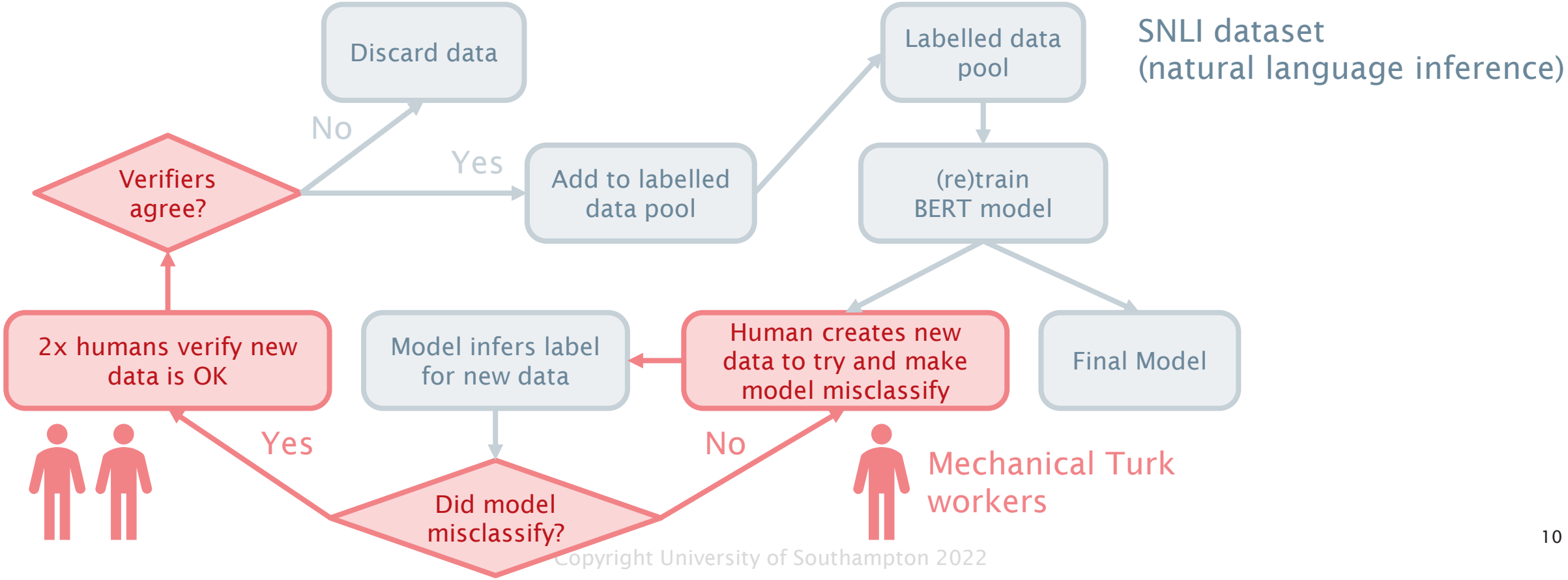
Learn accuracy of labelling function annotations from agreement/disagreement (no ground truth needed)

Learn to predict labels generalising beyond the labelling functions

University of Southampton

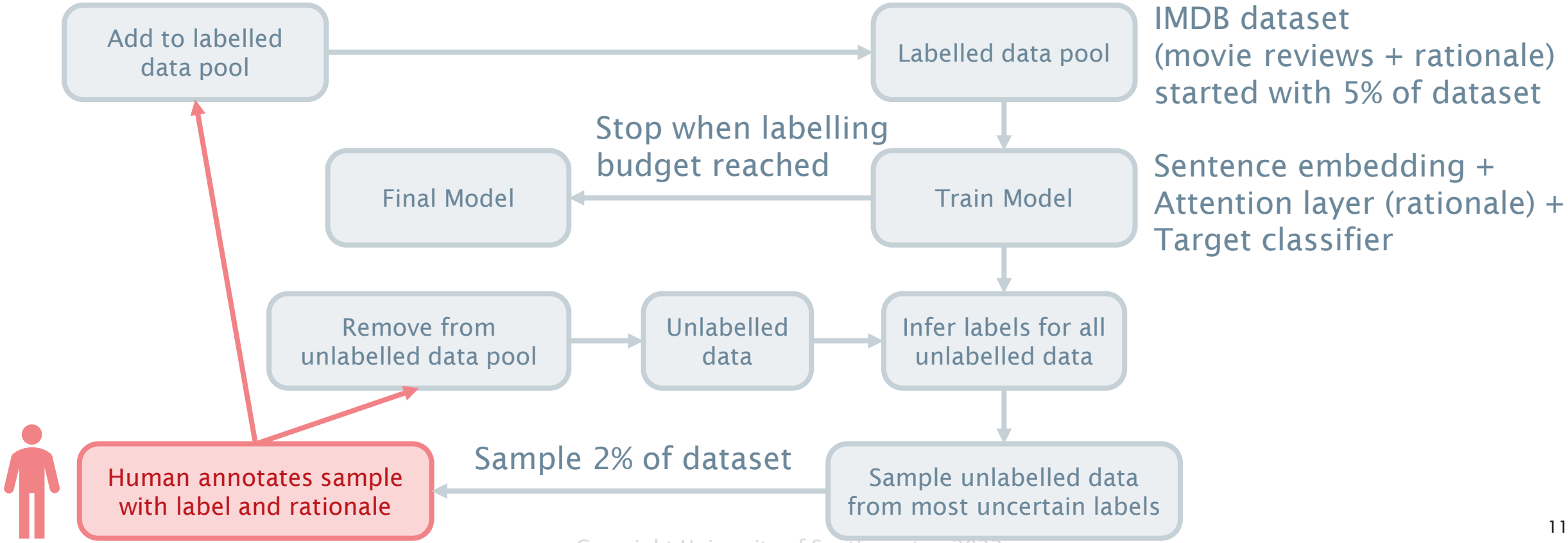
Human-in-the-loop NLP

- Adversarial training with human-in-the-loop [Nie 2020]
  - Humans creates novel examples that model has difficulty processing (like a GAN). This repeats in cycles, with model retrained each time to incrementally improving it



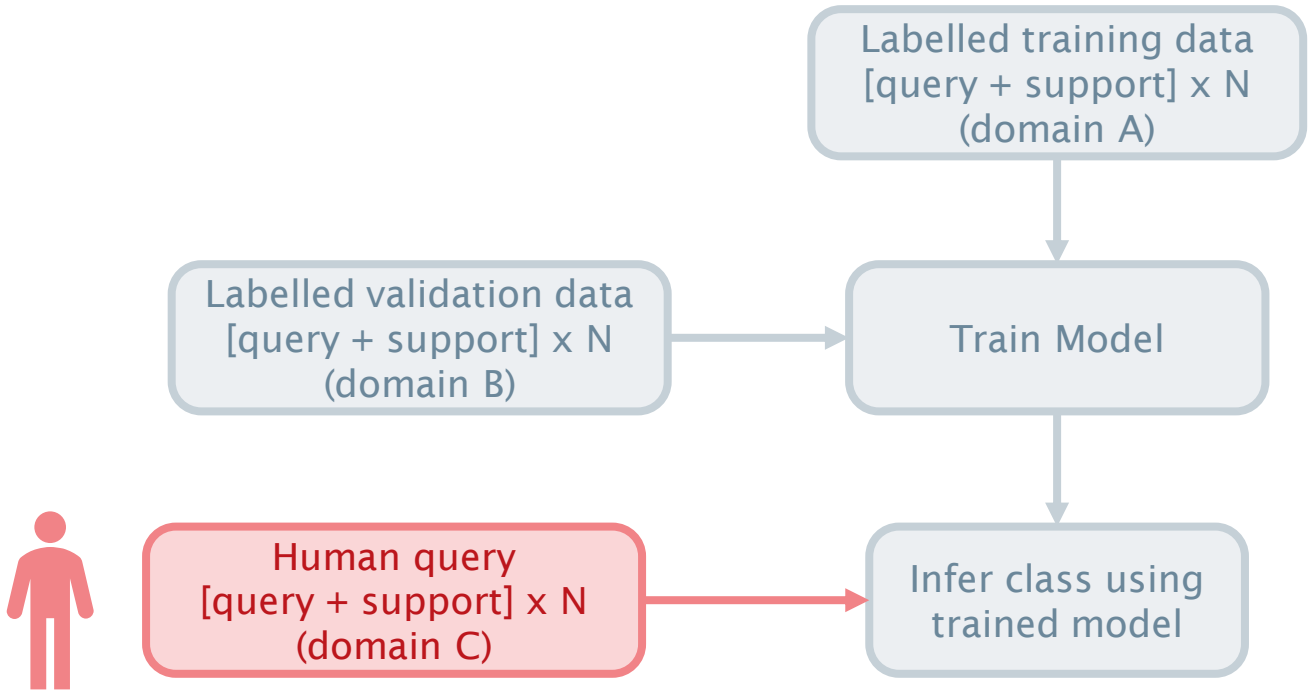
Human-in-the-loop NLP

- Active learning [Kanchinadam 2020]
  - Model chooses difficult unlabelled data for human to annotation, up to a max labelling budget. An attention layer is used to encode human rationale patterns also.



Human-in-the-loop NLP

- Few Shot NLP [Gao 2019]
  - Human provides an query item plus a support set and model tries to classify the query. FewRel 2.0 datasets tests domain adaption and None classes



FewRel 2.0 dataset
(relation extraction)
N-way = num of classes
K-shot = num instances of each class

Prediction = class of query sent
(relation type)

2-way 3 shot example
Query = sent X
Support = ((class 1, example sent 1, sent 2, sent 3), (class 2, example sent 1, sent 2, sent 3))

University of Southampton

Summary

- Human-in-the-loop AI (HILT)
  - Human in the training loop (offline)
  - Human in the deployment loop (online)
  - Humans can help AI models move beyond the limitations of static training data
- Human-in-the-loop NLP approaches
  - Interactive sense making
  - Explainable AI
  - Adversarial training
  - Active learning
  - Few Shot Learning and Meta-learning

Thank you for your attention!

Dr Stuart E. Middleton

University of Southampton, Electronics and Computer Science

email: sem03@soton.ac.uk

web: www.ecs.soton.ac.uk/people/sem03

twitter: @stuart_e_middle

This work was supported by the Engineering and Physical Sciences Research Council (EP/V00784X/1), Natural Environment Research Council (NE/S015604/1) and Economic and Social Research Council (ES/V011278/1).

University of Southampton

References

Bunch, E. You, Q. Fung, G. Human-In-The-Loop Topic Discovery with Embedded Text Representations. 1st Workshop on Data Science with Human in the Loop, DaSH@KDD 2020, 2020

Gao, T. Han, X. Zhu, H. Liu, Z. Li, P. Sun, M. Zhou, J. FewRel 2.0: Towards More Challenging Few-Shot Relation Classification, EMNLP-IJCNLP 2019

Ghai, B. Liao, Q.V. Zhang, Y. Mueller, K. Active Learning++: Incorporating Annotator's Rationale using Local Model Explanation, DaSH@KDD 2020, 2020

Hanafi, M.F. Abouzied, A. Danilevsky, M. Li, Y. WhyFlow: Explaining Errors in Data Flows Interactively. 1st Workshop on Data Science with Human in the Loop, DaSH@KDD 2020, 2020

Kanchinadam, T. Westpfahl, K. You, Q. Fung, G. Rationale-based Human-in-the-Loop via Supervised Attention, 1st Workshop on Data Science with Human in the Loop, DaSH@KDD 2020, 2020

Lee, H. Li, S. Vu, Y. Meta Learning for Natural Language Processing: A Survey. NAACL-2022

Middleton, S.E. Lavorgna, L. Neumann, G. Whitehead, D. Information Extraction from the Long Tail: A Socio-Technical AI Approach for Criminology Investigations into the Online Illegal Plant Trade, WebSci ’20 Companion, 2020

Middleton, S.E. Letouzé, E. Hossaini, A. Chapman, A. Trust, regulation, and human-in-the-loop AI: within the European region, Communications of the ACM (CACM), 65, 4 (April 2022), 64–68

Nie, Y. Williams, A. Dinan, E. Bansal, M. Weston, J. Kiela, D. Adversarial NLI: A New Benchmark for Natural Language Understanding, ACL 2020

University of Southampton

References

Ratner, A. Bach, S.H. Ehrenberg, H. Fries, J. Wu, S. Ré, C. Snorkel: rapid training data creation with weak supervision. Proc. VLDB Endow. 11, 3 (November 2017), 269–282

Yin, W. Meta-learning for Few-shot Natural Language Processing: A Survey, arXiv,2020, https://doi.org/10.48550/arXiv.2007.09604