University of BRISTOL

Aggregating Weak Annotations from Crowds

Edwin Simpson

Department of Computer Science,

University of Bristol, UK.

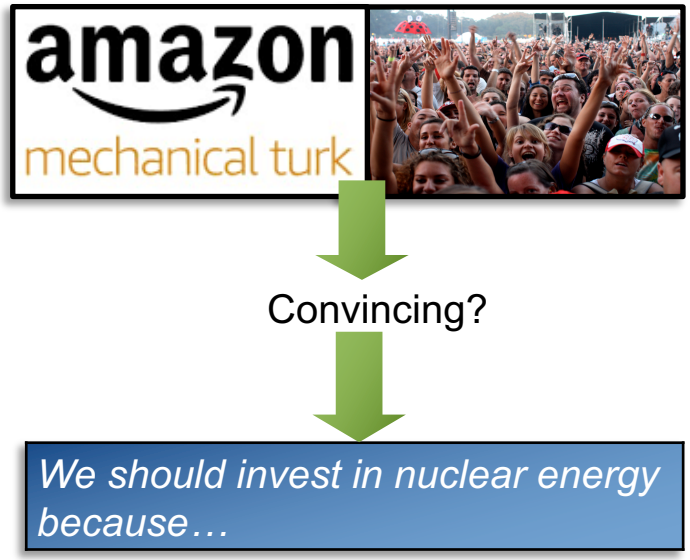
bristol.ac.uk

UBIQUITOUS KNOWLEDGE PROCESSING

TECHNISCHE UNIVERSITÄT DARMSTADT

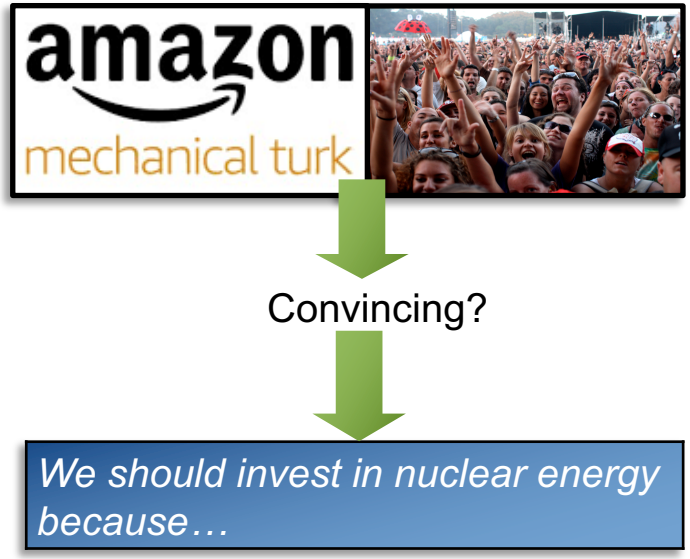
Crowdsourcing Training Data

- We need training and evaluation data for machine learning methods for new domains, users and tasks.
- Crowdsourcing offers a cheap, fast solution by paying non-expert annotators a few cents per annotation;
- Requires tasks that can be broken into small, easy steps with clear instructions.



Crowds Provide Weak Labels

- Many apparent errors caused by various sources of **disagreement**:
  - Mistakes
  - Spammers
  - Ambiguity
  - Subjectivity.
- Labels from multiple annotators must be aggregated to estimate a ‘gold’ standard.
- How can we deal with these types of disagreement when aggregating weak annotations from crowds?



In This Talk:

- Part 1: Preference learning for ambiguous annotation tasks
- Part 2: Aggregating crowdsourced class labels

Part 1. Preference Learning for Ambiguous Annotation Tasks

How Convincing is this Argument?

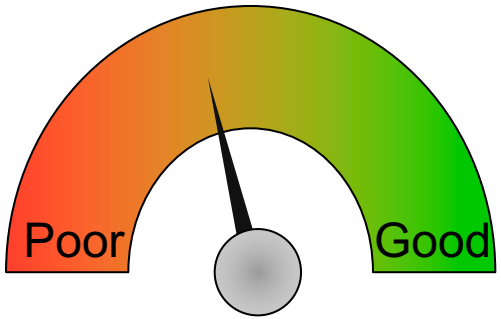
- Faced with a large text corpus
  - Social media, historical archives…
  - Find me the most persuasive arguments on topic X;
  - E.g. investing in nuclear energy.
- Goal: learn and predict convincingness
- Convincing vs. not convincing – where is the cut-off?

Finding Convincing Arguments using Scalable Bayesian Preference Learning, Simpson & Gurevych, TACL, 2018.

How Convincing Is This Argument?

- Numerical score
  - ✘ Annotators are inconsistent over time
  - ✘ Annotators interpret scores differently: Bob's 4 stars == Alice's 5 stars
  - ✘ Often have a fixed number of categories (e.g., integers from 1 to 5)
- Example topic: Firefox is better than Internet Explorer:

Who said anything about FF/mozilla having anything to do with steve jobs/apple?
Seperate enititys, not in the least bit attached!
BTW, IE blows when making pages and…



Which Argument is More Convincing?

- Numerical or categorical scoring
  - ✘ Annotators are inconsistent over time
  - ✘ Annotators interpret scores differently: Bob's 4 stars == Alice's 5 stars
  - ✘ Often have a fixed number of categories (e.g., integers from 1 to 5)
- Pairwise preferences:
  - ✓ Greater precision -- total sorting
  - ✓ Can be faster to label and reduce cost
  - ✓ No calibration needed for different annotators

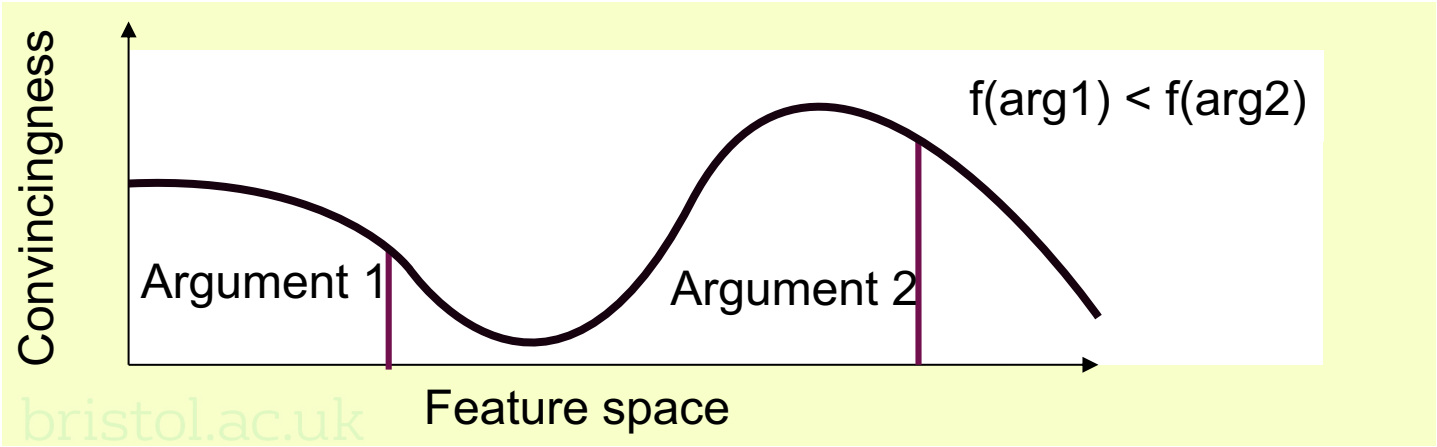
Who said anything about FF/mozilla having anything to do with steve jobs/apple? Seperate enititys, not in the least bit attached! BTW, IE blows when making pages and …	Firefox takes the best of all previous browsers and sticks it all in one neat package. Security and extendibility are some of its top features. And those times when …

A Model for Learning from Pairwise Preferences

- Goal: predict degrees of convincingness from pairwise labels

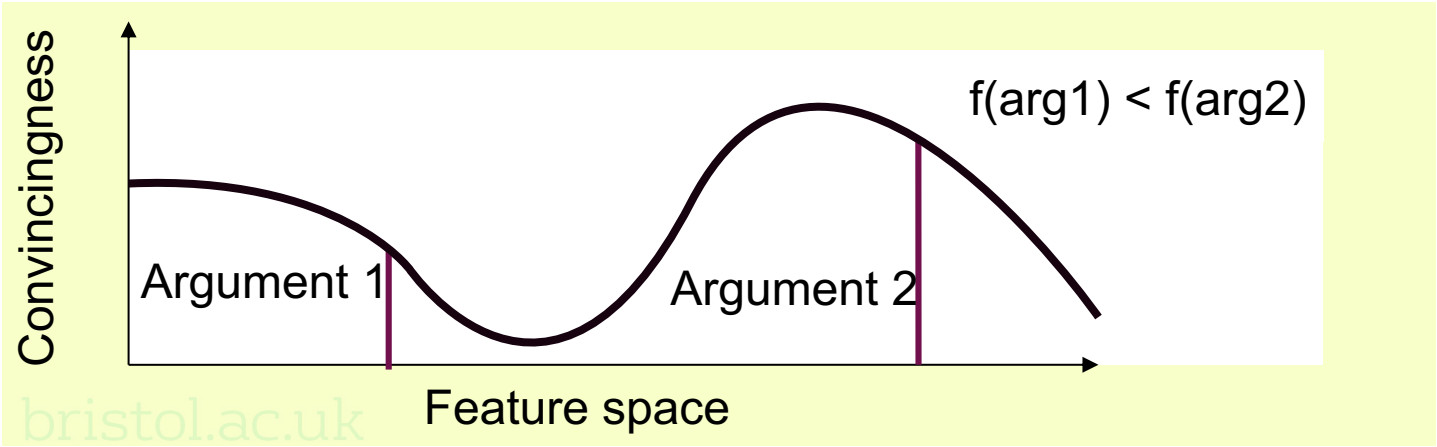
A Model for Learning from Pairwise Preferences

- Goal: predict degrees of convincingness from pairwise labels



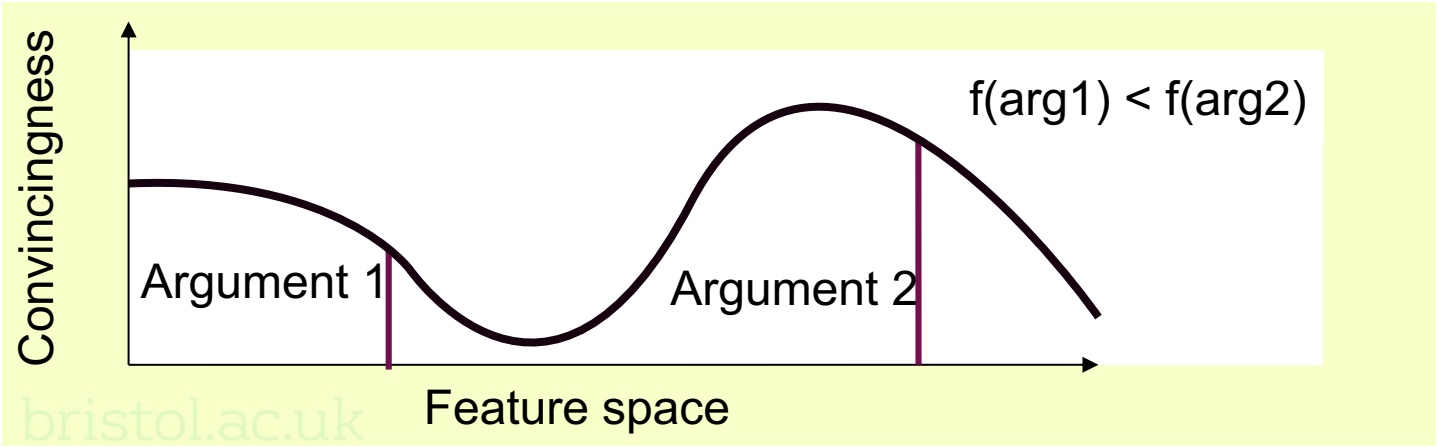
A Model for Learning from Pairwise Preferences

- Goal: predict degrees of convincingness from pairwise labels
- Challenges:
  - Annotation errors and disagreements between annotators
  - Sparse data – new domains and tasks, limited labelling budget



A Model for Learning from Pairwise Preferences

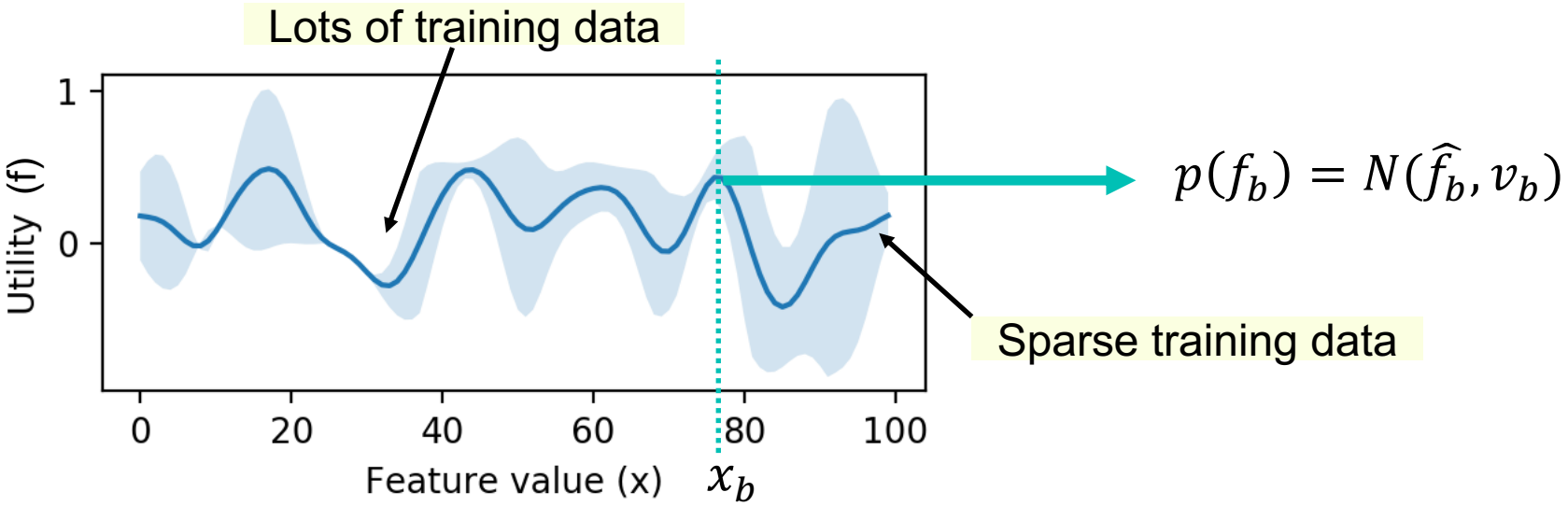
- Goal: predict degrees of convincingness from pairwise labels
- Challenges:
  - Annotation errors and disagreements between annotators
  - Sparse data – new domains and tasks, limited labelling budget



Bayesian solution?

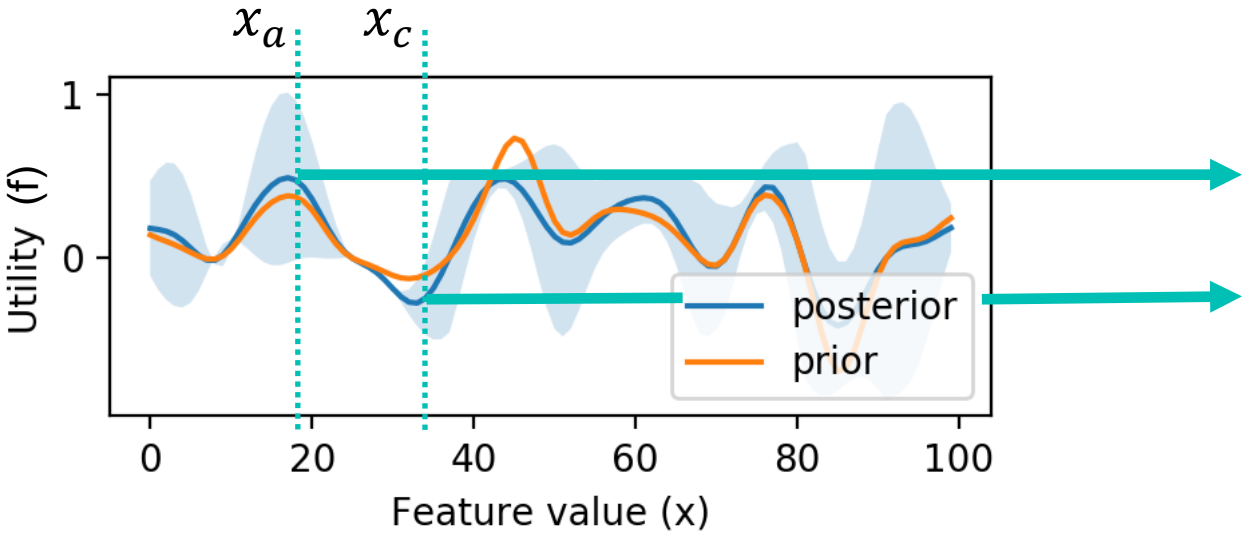
Gaussian Process Preference Learning (GPPL)

- A Gaussian Process (GP) is a probability distribution over functions
  - Posterior provides confidence estimates (variance)
  - Especially useful where we have little data



Gaussian Process Preference Learning (GPPL)

- Chu & Ghahramani (2005) introduced GP preference learning
  - Infer a posterior over functions given pairwise preferences
  - Simpson & Gurevych (2018) proposed a scalable variant



Pairwise observation:
$a \succ c$

Preference Learning Experiments

Argument Convincingness

- 1,052 arguments from createdebate.com & procon.org
- Amazon MTurkers were asked: ”Which is more convincing?”
- ~17,000 pairs with 5 annotations each for 32 topics
- Gold standard produced by redundant labelling + removing contradictory preferences

Who said anything about FF/mozilla having anything to do with steve jobs/apple?...	Firefox takes the best of all previous browsers and sticks it all in one neat package….

Data from: *Which argument is more convincing?* Habernal & Gurevych, ACL 2016.

Training on Crowdsourced Preferences

- Train on pairs for 31 topics, test on the remaining topic
  - Training data: one crowdsourced label per pair (no data cleaning)
  - Contains noise and conflicting preferences
  - Two tasks: predict pairwise labels, rank arguments
- GP inputs: linguistic features + mean GloVe embeddings

Convincingness Results

- GPPL learns a single model for both tasks directly from pairs
  - Others do regression on the gold convincingness scores in training topics
  - GP gives better handling of noise and data sparsity

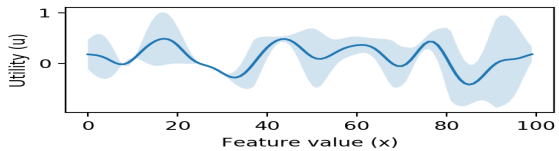
Method	Pair Prediction Accuracy	Rank correlation Kendall
SVM	.70	.31
BiLSTM	.73	.21
GPPL	**.77**	**.40**

Active Learning Exploits Uncertainty

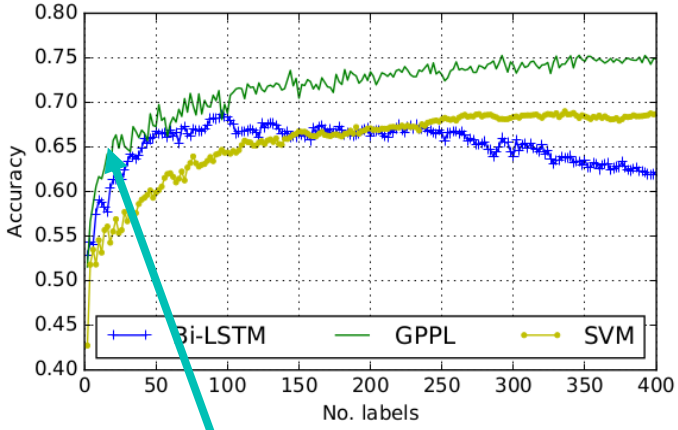
- Goal: reduce number of annotations
- Label only the most informative documents
- Simulation:
  1. Choose a random seed of 2 pairs
  2. Train the model
  3. Evaluate accuracy
  4. Get labels for 2 most uncertain document pairs
  5. Repeat from step 2

Active Learning Exploits Uncertainty

- Goal: reduce number of annotations
- Label only the most informative documents
- Simulation:
  1. Choose a random seed of 2 pairs
  2. Train the model
  3. Evaluate accuracy
  4. Get labels for 2 most uncertain document pairs
  5. Repeat from step 2



Argument convincingness:



GPPL learns rapidly with small numbers of labels

Acquiring Weak Labels from End Users

Simpson, E., Gao, Y., & Gurevych, I. (2019). Interactive Text Ranking with Bayesian Optimisation: A Case Study on Community QA and Summarisation. *arXiv preprint arXiv:1911.10183*.

Motivation for Acquiring User Feedback

- Many tasks are **highly specific** to a given topic or user
- E.g. community question answering:
  - Think of StackExchange, Quora
  - Questions describe an individual user's problem
- E.g. summarisation:
  - Documents to summarise discuss a narrow topic
- Generic models may have issues meeting a user's needs:
  - Users may under-specify what they want when writing a question
  - Domain knowledge is required
- Hence, text ranking tasks often benefit from **user feedback**

Interactive Text Ranking

Acquisition function

Query User

Which summary do you prefer?

☐ **(b):** The arrests of Xu and Qin at their homes Monday night and the accusations against them and Wang were the sharpest action Chinese leaders have taken since dissidents began pushing to set up and legally register the China Democracy Party in June. Hours before China was expected to sign a key U.N. human rights treaty and host British Prime Minister Tony Blair, police hauled a prominent human rights campaigner in for questioning Monday. With attorneys locked up, harassed or plain scared, two prominent dissidents will defend themselves against charges of subversion Thursday in China's highest-profile dissident trials in two years. Wang was a student leader in the 1989 Tiananmen Square democracy demonstrations.

☑ **(c):** On the eve of China's signing the International Covenant of Civil and Political Rights (ICCPR) in October 1998, police detained Chinese human rights advocate Qin Yongmin for questioning. Eight weeks after signing the ICCPR, Chinese police arrested Qin and an associate in the China Democracy Party (CDP), Xu Wenli, without stating charges. Another CDP leader already in custody, Wang Youcai, was accused of "inciting the overthrow of the government". Qin and Wang went to trial in December for inciting subversion. Police pressure on potential defense attorneys forced the accused to mount their own defenses. Xu Wenli had not yet been charged.

Candidate Pool

Predict utilities

Ranking Model

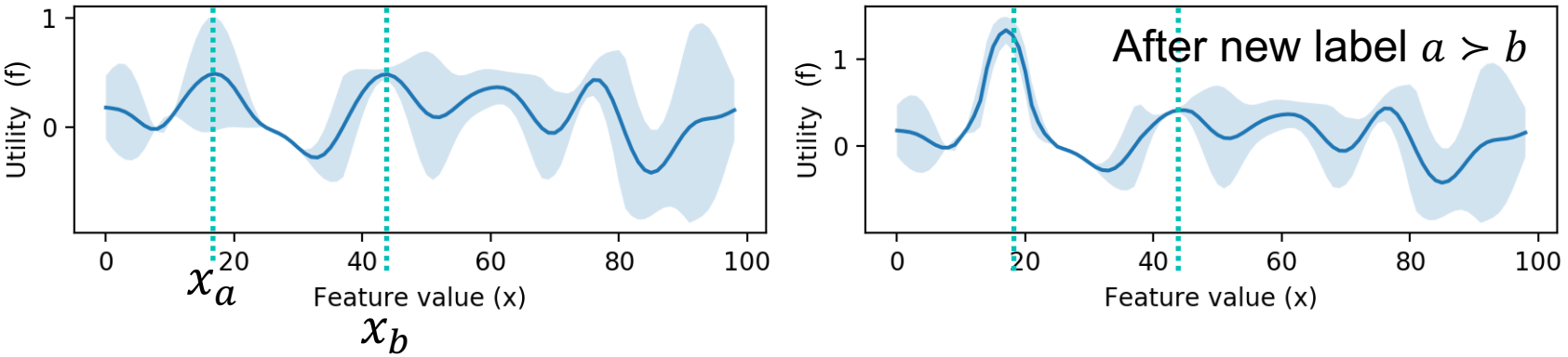
Retrain

Labels

bristol.ac.uk

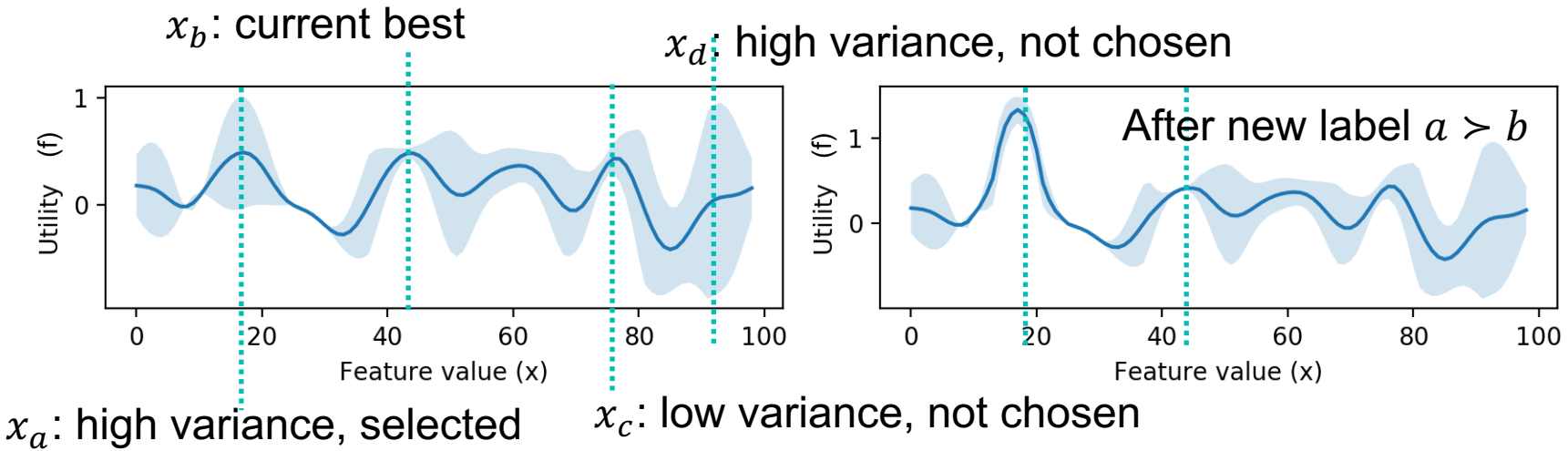
Bayesian Optimisation

- Goal: find the best candidate with as few user labels as possible
- Improvement $= \max\{f_a - f_b, 0\}$



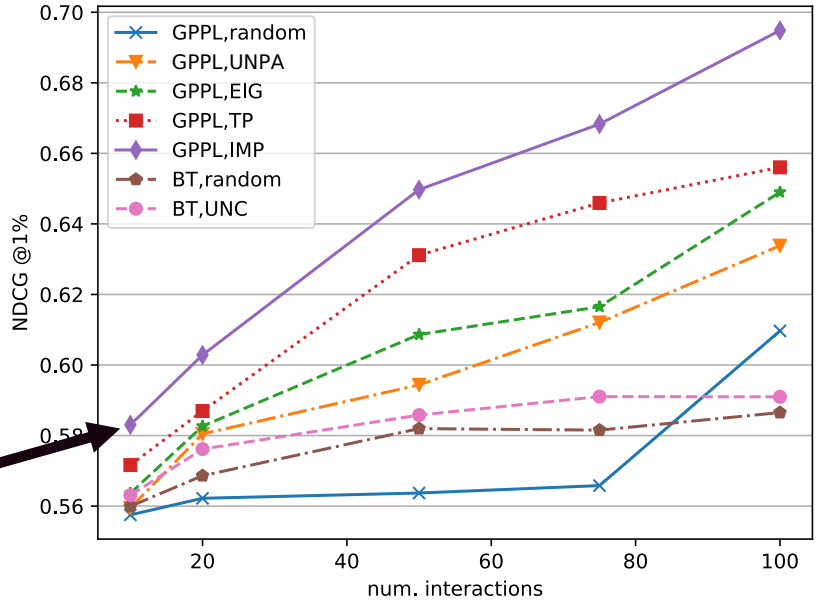
Bayesian Optimisation

- $\mathrm{IMP}(a, \boldsymbol{D}) = \mathbb{E}[\max\{f_a - f_b, 0\}]$
- Selected pair = current best *b,* candidate *a* that maximises IMP
- Collect labels that improve the final result we show to the user



Interactive Summarisation

- Extractive multi-document summarisation
- Ranking performance on DUC'01 dataset →
- Relevance of 100 top-ranked summaries
- GPPL improves over linear models
- Bayesian optimisation learns quickest (GPPL,IMP)



Results

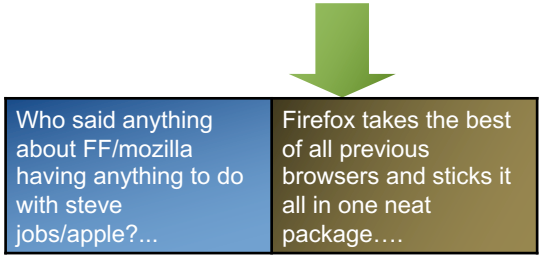
	cQA (accuracy)	Multi-document summarisation (mean combined ROUGE)
	StackExchange	DUC newswire data
	10 interactions	20 interactions
No interactions	0.44	1.82
UNC (uncertainty sampling)	0.30	1.83
EIG (expected information gain)	0.38	1.90
IMP (BO expected improvement)	**0.72**	**2.03**

Please see Interactive Text Ranking with Bayesian Optimisation: A Case Study on Community QA and Summarisation for lots more results and discussion

Aggregating Preferences from Biased Individuals

Adapting GPPL To Cope With Subjectivity

- Each user provides a limited number of pairwise annotations;
- Predict **preferences of individual users** in a crowd;
- Use the model of individual biases to improve our estimate of the **consensus;**
- Exploit similarities between users' preferences to address sparsity (collaborative filtering);



Simpson, E., & Gurevych, I. (2020). Scalable Bayesian preference learning for crowds. *Machine Learning*, 1-30.

CrowdGPPL

- Assume a *consensus* preference function, modelled by GPPL
- Assume multiple *latent factors*, each modelled by GPPL
- Each *latent factor* captures an interest shared by multiple users
- Each individual's preferences are a weighted combination of latent factors and the consensus

$$f(\boldsymbol{x}_a, \boldsymbol{u}_j) = \sum_{c=1}^{C} v_c(\boldsymbol{x}_a) w_c(\boldsymbol{u}_j) + t(\boldsymbol{x}_a)$$

personal rating

latent factor rating

x user weight for latent factor

+ consensus rating

Preference Learning with Crowds: CrowdGPPL

- CrowdGPPL provdes a model for estimating both personalised and consensus ratings.
- A new, scalable method for Bayesian matrix factorisation using stochastic variational inference, Hoffman et al. (2013).
- Predicting consensus preferences for argument convincingness:
- Improves predictions by accounting for individual preferences and labelling error rates.

Method	Pair Prediction Accuracy	Rank correlation Kendall
GPPL	.77	.50
CrowdGPPL	**.79**	**.53**

Part 2. Aggregating Crowdsourced Class Labels

The Power of Crowdsourcing

- Assume annotators make uncorrelated, zero-mean errors
- Average squared error of an individual:

$$e_{av} = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\epsilon_k(x)^2]$$

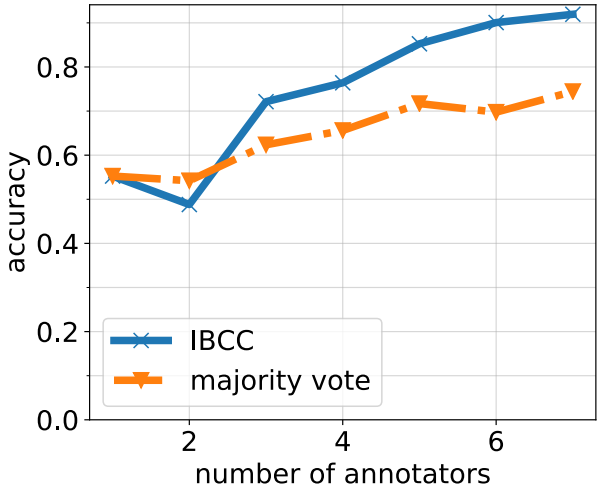
- Expected squared error of a combination:

$$e_{com} = \mathbb{E}[\frac{1}{K}\sum_{k=1}^{K}\epsilon_k(x)^2] = \frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}[\epsilon_k(x)^2] = \frac{1}{K}e_{av}$$

- So the more crowdworkers, the better*.
- *if the assumptions stick😉

It Helps to Model Individuals' Errors…

- Not all annotators are equal!
  - Spammers, guessers
  - Different abilities
  - Boredom/enthusiasm
  - …
- Types of error:
  - Noise or variance = 'random' errors, i.e. we don't know the reason for the error
  - Bias = consistently choosing certain incorrect labels, e.g., spamming



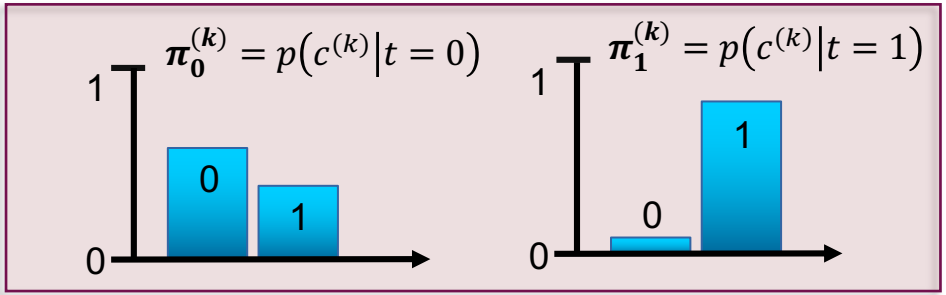
Performance on Simulated Data

IBCC: Independent Bayesian Classifier Combination

- A probabilistic, generative model, based on Dawid & Skene (1979) and given Bayesian treatment by Kim & Ghahramani (2012)
- Given a true label, *t,* we observe noisy labels, ***c,*** from *K* annotators

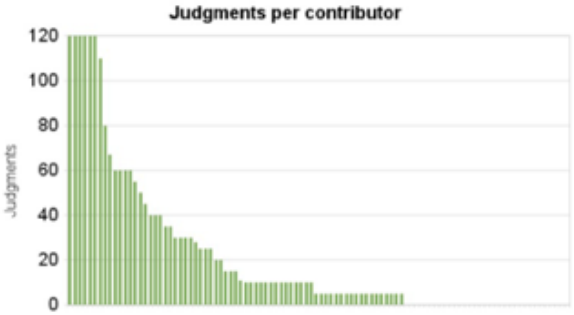
IBCC: Independent Bayesian Classifier Combination

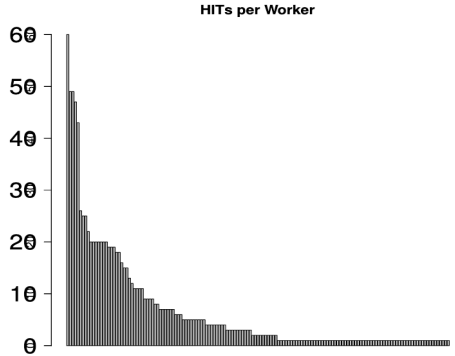
- A probabilistic, generative model, based on Dawid & Skene (1979) and given Bayesian treatment by Kim & Ghahramani (2012)
- Given a true label, *t,* we observe noisy labels, ***c,*** from *K* annotators
- Likelihood $\boldsymbol{\pi}_j^{(k)}$ of an annotation captures noise and bias



Bayesian Inference for IBCC

- Infer posterior $p(t = j|\boldsymbol{c})$ and likelihoods $\boldsymbol{\pi}_{\boldsymbol{j}}^{(\boldsymbol{k})}$
- Overfitting with maximum likelihood
- Most workers only do a few tasks
- Do you trust the person who did 5 questions and got them all right…
- ...as much as the person who got 90% of 1000?

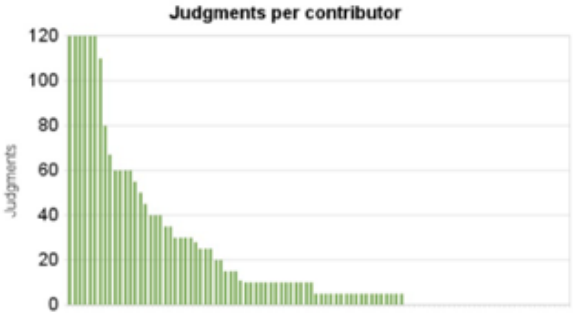


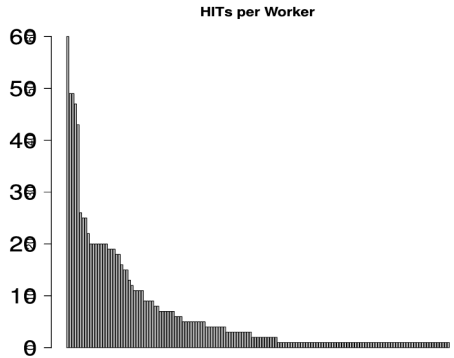


Combined Decision Making with Multiple Agents, Simpson, DPhil Thesis, 2014.
Ground truth generation in medical imaging: a crowdsourcing-based iterative approach, Rodriguez et al. (2012).
Crowdsourcing Document Relevance Assessment with Mechanical Turk, Grady and Lease (2010).

Bayesian Inference for IBCC

- Infer posterior $p(t = j|\boldsymbol{c})$ and likelihoods $\boldsymbol{\pi}_{\boldsymbol{j}}^{(\boldsymbol{k})}$
- Bayesian inference:
  - Accounts for uncertainty in model parameters
  - Confidence estimates with 'small' and noisy data
- Variational Bayes
  - Approximates a Bayesian solution with much more rapid convergence



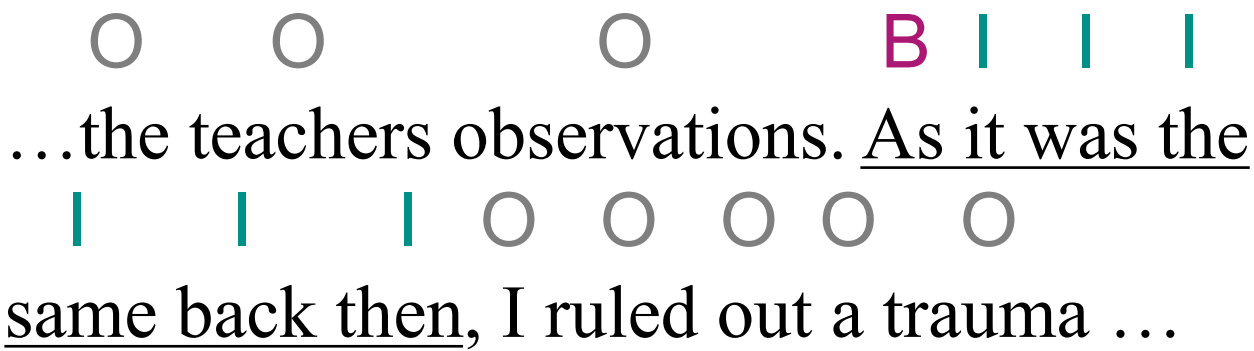


Combined Decision Making with Multiple Agents, Simpson, DPhil Thesis, 2014.
Ground truth generation in medical imaging: a crowdsourcing-based iterative approach, Rodriguez et al. (2012).
Crowdsourcing Document Relevance Assessment with Mechanical Turk, Grady and Lease (2010).

Sequence Labelling with Crowds

A Better Model for Sequence Labelling?

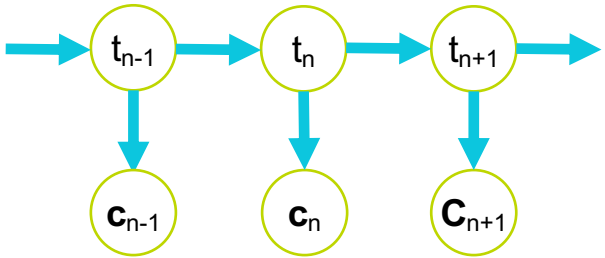
- In text annotation schemes such as IOB2, the probability of a label depends on what came before it.
- E.g. ‘Inside’ (I) cannot follow ‘Outside’ (O):



A Bayesian Approach for Sequence Tagging with Crowds, Simpson & Gurevych, EMNLP 2019.

A Better Model for Sequence Labelling?

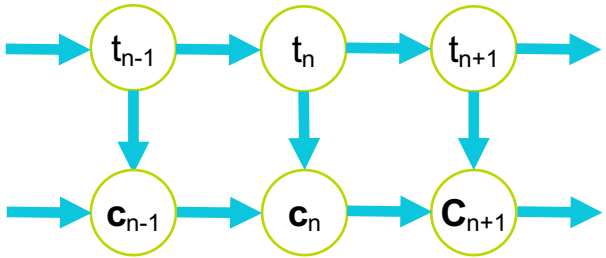
- **Bayesian sequence combination (BSC)**:
  - True label depends on previous true label: Hidden Markov Model (HMM)
  - Tokens and annotations as observations



A Bayesian Approach for Sequence Tagging with Crowds, Simpson & Gurevych, EMNLP 2019.

A Better Model for Sequence Labelling?

- **Bayesian sequence combination (BSC)**:
  - True label depends on previous true label: Hidden Markov Model (HMM)
  - Tokens and annotations as observations
  - Annotation likelihood $\boldsymbol{\pi}_j^{(k)}$ depends on the annotator's previous label

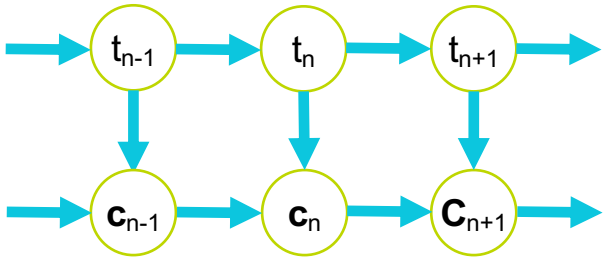


A Bayesian Approach for Sequence Tagging with Crowds, Simpson & Gurevych, EMNLP 2019.

A Better Model for Sequence Labelling?

- **Bayesian sequence combination (BSC)**:
  - True label depends on previous true label: Hidden Markov Model (HMM)
  - Tokens and annotations as observations
  - Annotation likelihood $\boldsymbol{\pi}_j^{(k)}$ depends on the annotator's previous label

✓ Recognises that some transitions are illegal
✓ Models bias toward spans that are too short or too long
✗ Number of parameters = num_classes^3



A Bayesian Approach for Sequence Tagging with Crowds, Simpson & Gurevych, EMNLP 2019.

BSC: Experiments with Crowdsourcing

- Named entity recognition, CoNLL 2003 (NER)
- Medical trial populations (PICO)
- Argument span identification (ARG)

BSC: Experiments with Crowdsourcing

- Named entity recognition, CoNLL 2003 (NER)
- Medical trial populations (PICO)
- Argument span identification (ARG)

F1 scores	NER	PICO	ARG
Best worker	67.3	58.5	60.0
Majority vote	65.4	64.3	34.8

BSC: Experiments with Crowdsourcing

- Named entity recognition, CoNLL 2003 (NER)
- Medical trial populations (PICO)
- Argument span identification (ARG)

- Modelling annotators helps:

F1 scores	NER	PICO	ARG
Best worker	67.3	58.5	60.0
Majority vote	65.4	64.3	34.8
MACE	70.0	39.0	32.0
IBCC	74.4	68.9	46.4

BSC: Experiments with Crowdsourcing

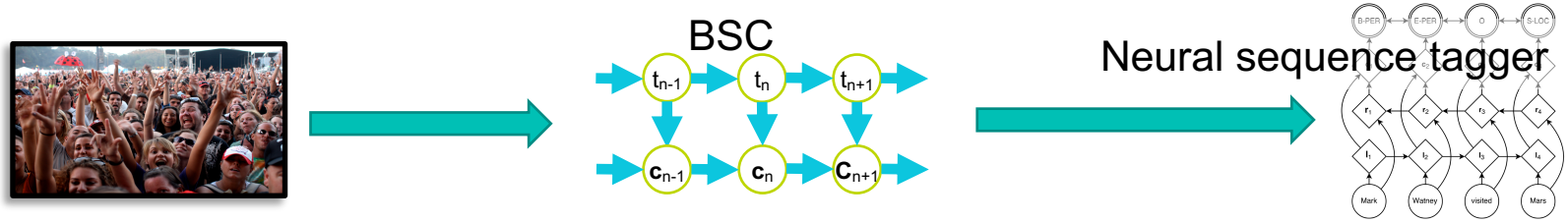
- Named entity recognition, CoNLL 2003 (NER)
- Medical trial populations (PICO)
- Argument span identification (ARG)

F1 scores	NER	PICO	ARG
Best worker	67.3	58.5	60.0
Majority vote	65.4	64.3	34.8
MACE	70.0	39.0	32.0
IBCC	74.4	68.9	46.4
HMM-crowd	74.2	71.0	40.0
BSC	**77.4**	**72.8**	**60.1**

- Modelling sequential dependencies improves aggregated labels:

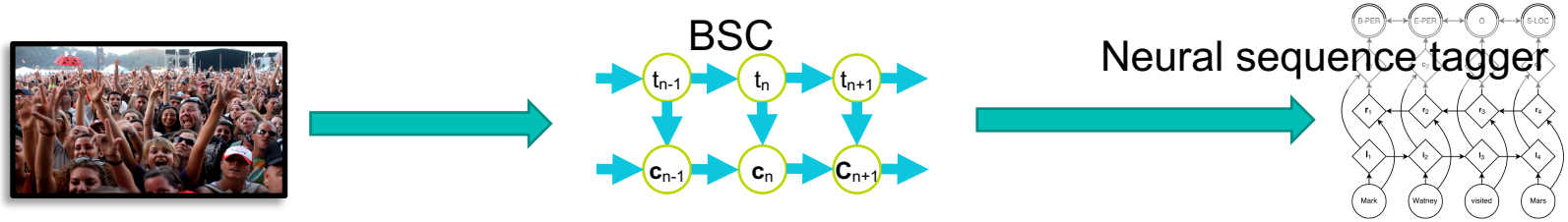
Nguyen et al. (2017) Aggregating and predicting sequence labels from crowd annotations

The Aggregation and Training Pipeline



The Aggregation and Training Pipeline

- BSC outputs a “gold standard” for training a sequence tagger.

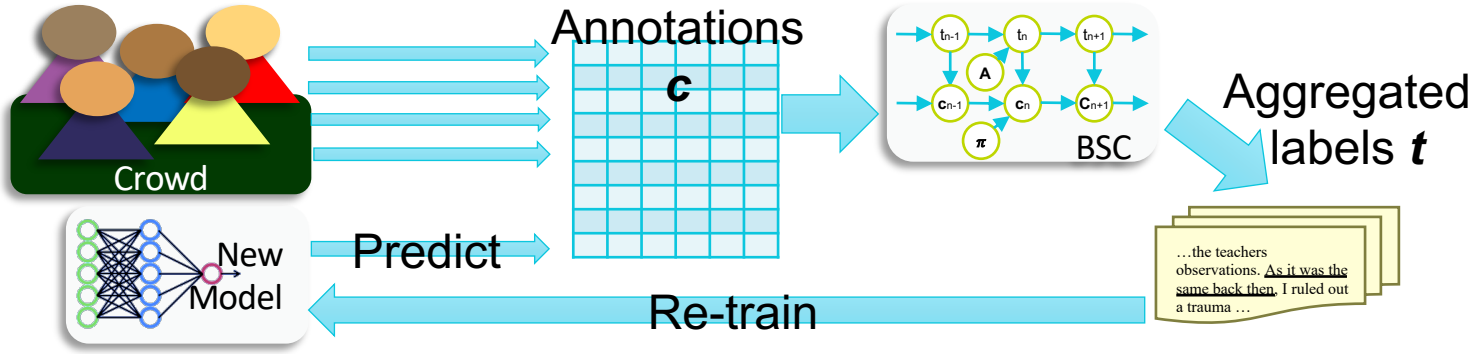


- BSC models label sequences using a hidden Markov model (HMM):
  - Training an HMM from scratch is often highly sub-optimal;
  - LSTMs or transformers perform much better thanks to representation learning;
  - Transformers pretrain embedding representations on huge unlabelled datasets.
- Couple BSC with a sequence tagger to learn directly from crowd labels:
  - Make use of specialised sequence taggers to improve the aggregation.
  - Indicate disagreement and confidence to the sequence tagger.

Using a Sequence Tagger During Aggregation

Several ways we can produce the gold standard from crowdsourced labels:

1. BSC.
2. Pipeline: BSC → sequence tagger.
3. Variational combined supervision (VCS): integrated sequence tagger.



Low Resource Sequence Tagging with Weak Labels, Simpson, Pfeiffer & Gurevych, AAAI 2020.

Variational Combined Supervision

- Named entity recognition, CoNLL 2003 (NER)
- Medical trial populations (PICO)
- Sequence tagger: BiLSTM-LSTM-CRF (Lample et al., 2016)

- The integrated sequence tagger improves performance of the aggregator.
- Could integrate any specialised model for a particular domain.

F1 scores	NER	PICO
BSC	**77.4**	**72.8**
Pipeline	77.7	75.5
BSC with VCS	78.0	77.5

Low Resource Sequence Tagging with Weak Labels, Simpson, Pfeiffer & Gurevych, AAAI 2020.

Conclusions

- Practical machine learning problems call for acquiring training data from multiple annotators, e.g., by crowdsourcing;
- Design annotation tasks that reduce ambiguity: think preference learning!
- Handle errors and disagreements in the annotations using Bayesian methods – much more label-efficient than majority vote;
- Future: Bayesian variant of crowd layers (e.g., Peters et al., (2018), Rodrigues et al., (2018))?
- Future: evaluate models for different kinds of data, especially structured outputs and numerical ratings.

Discussion…

- Software and methods are available to reuse:
  - GPPL and CrowdGPPL: preference learning from crowds
  - IBCC and BSC: aggregate classifications/sequence labels
- Joint work alongside Iryna Gurevych, Erik-Lân Do Dinh, Tristan Miller, Jonas Pfeiffer, Yang Gao, Ivan Habernal
- Discussion points:
  - New, richer modes of user interaction
  - How to make crowdsourcing a less manually-intensive process
  - Flaws in assuming a single ‘gold’ standard