figure eight FEDERAL

Figure Eight Federal and Appen Capabilities

HIATUS Proposer's Day

Presented by: Jefferson Barlew, Senior Linguist, Advanced Research and Solutions

January 19, 2022

Who is Figure Eight Federal and Appen

The Figure Eight Federal (F8F) and Appen team, has a long history associated with NLP. Our team has strengths in leveraging expert linguist and native informant capabilities for complex dataset design and data collection, transcription, translation, annotation and evaluation activities in 100+ languages, including many languages that are low-resource, typologically diverse and geo-politically significant. This proven past performance and technical acumen are foundational elements which will be leveraged for this program.

- Our experience has been used to advance techniques for evaluating both the sense and soundness in machine-modified informational texts as well as the human-interpretable explanations for NLM text classifiers.

Capability: Dataset Provision

Experience

IARPA

Babel

MATERIAL

DARPA

AIDA

KAIROS

Collaboration within and across performer teams

Expertise

Market Leader in ML training data

Expert linguists to provide QA, annotation, analysis, etc.

Extensive experience across the entire AI lifecycle from training to deployment and evaluation

Crowd

> 1,000,000 native speakers

Hundreds of languages

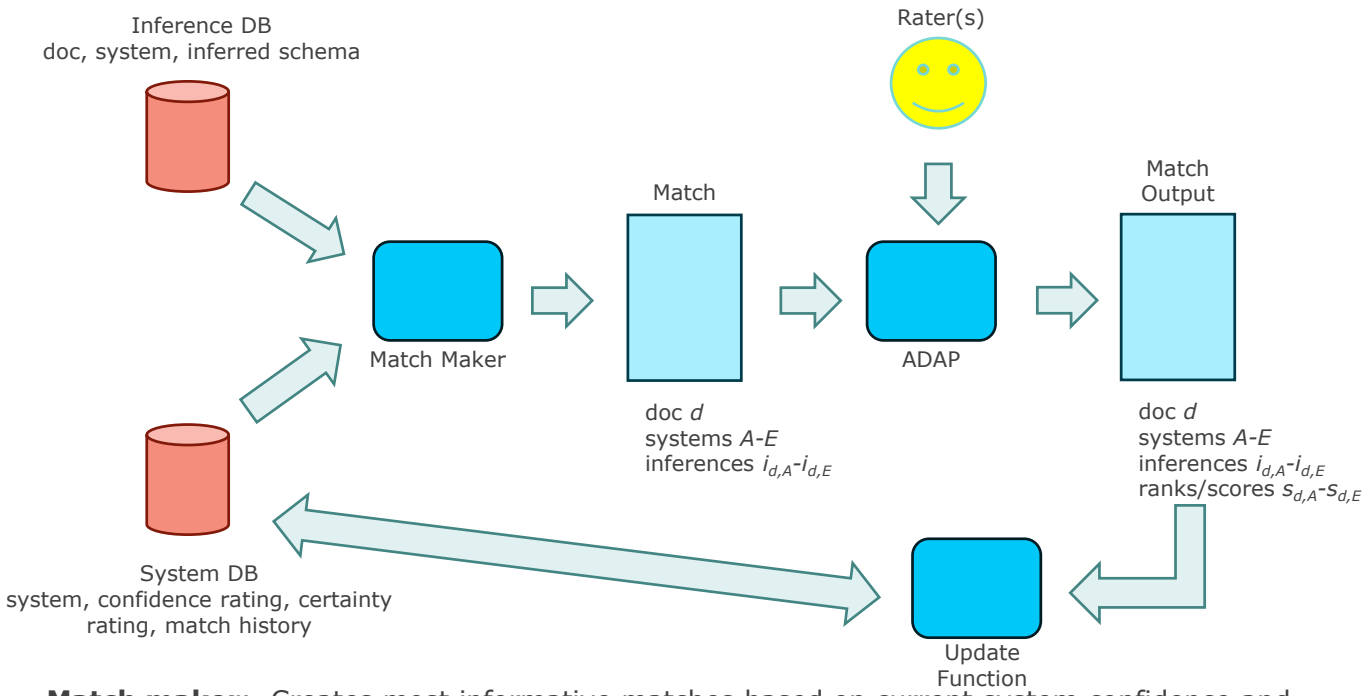
Typological diversity

Geo-political significance

Capability: Model Evaluation

- Evaluates multiple models simultaneously
- Uses comparative and absolute rankings
- Reduces data requirements with active learning
- Quickly reaches significant results
- Connects to Appen's modular annotation ecosystem
- Leverages relatively low-cost contributor pools
- Developed for and tested in DARPA KAIROS
- Implements EASL methodology (Sakaguchi et al. 2014, Sakaguchi and Van Durme 2018).

Model Evaluation Example: Pipeline and Results

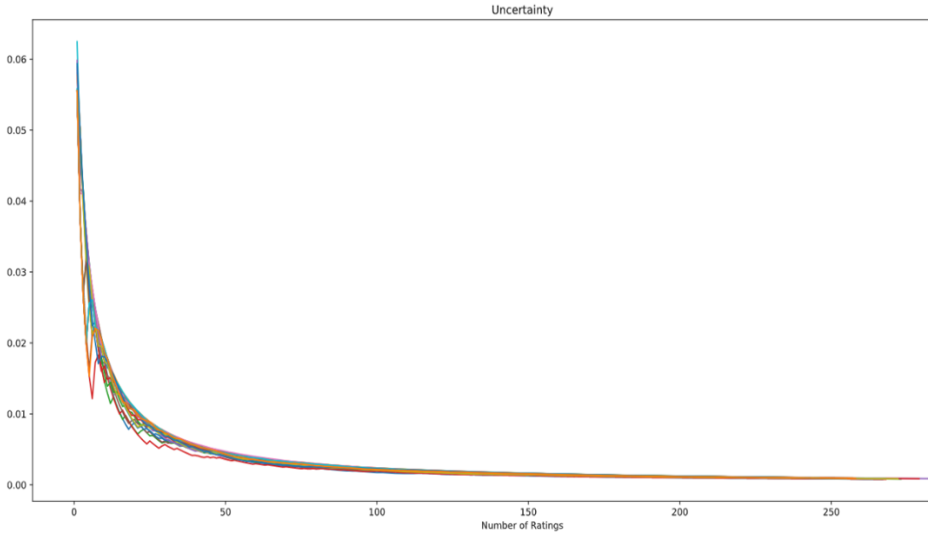


Match maker: Creates most informative matches based on current system confidence and uncertainty ratings.

Appen Data Annotation Platform (ADAP): Presents a rater with a single doc *d* plus inferences $i_{d,A}$-$i_{d,E}$ based on that doc and made by 5 different systems. Rater inputs rank + absolute score for each inference.

Update Function: Takes the match output and updates confidence and uncertainty ratings for participating systems

Results



Data: Triples extracted from unstructured data by ~25 models

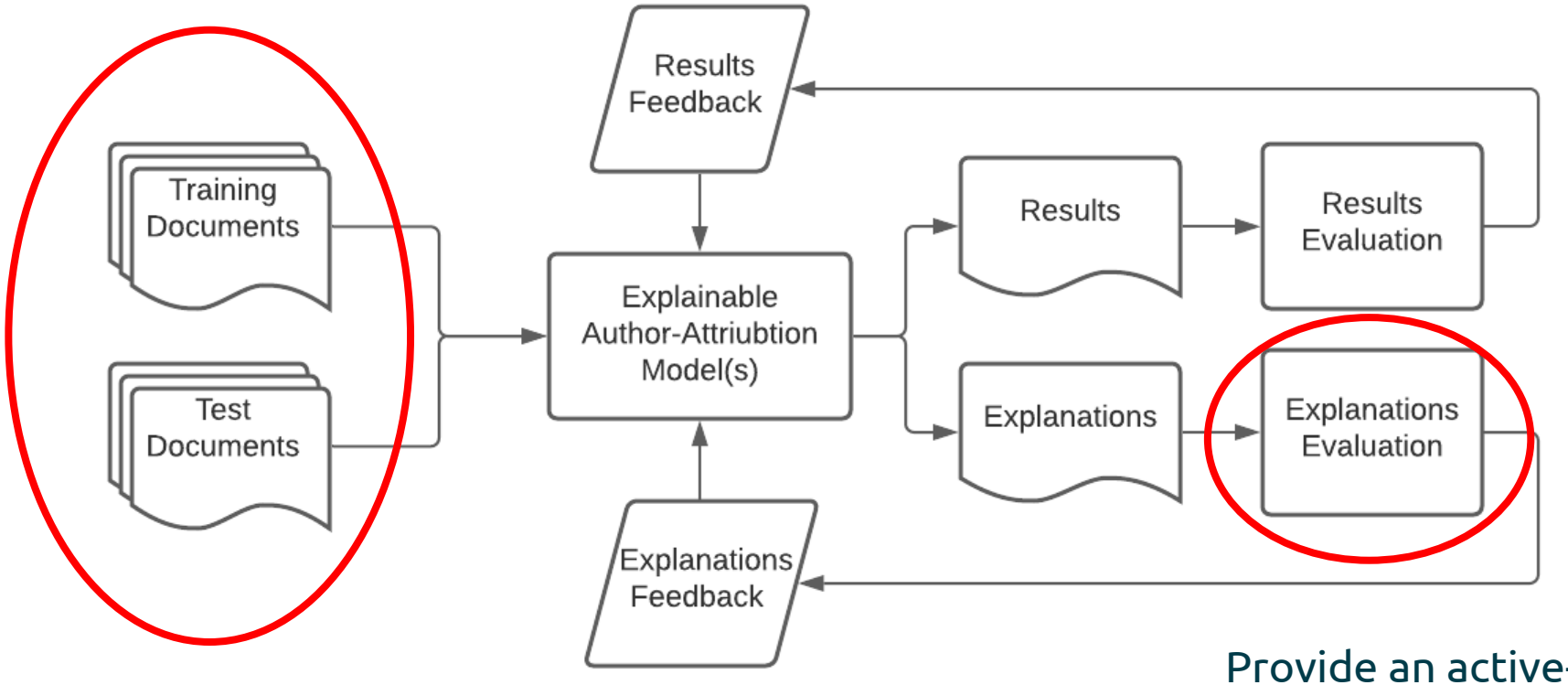
Evaluation: Naïve human contributors evaluate 5 extractions at a time for naturalness and plausibility using a scaled slider

Y-axis: The uncertainty score for the quality of each model, where low uncertainty corresponds to high reliability for the quality score (quality scores not pictured)

X-axis: The number of extractions evaluated from each model

Takeaway: By around 50 annotations per model, uncertainty is low and reliability is high for all models, making evaluation of additional data unnecessary and yielding informative results to model-builders quickly.

Notional Explainable Author Attribution Pipeline



Assumptions

- Integrated explainable models or separate explanation generating models
- Multiple models across and within performer teams that require comparative evaluations
- Author attribution results evaluated primarily by gold-standard data

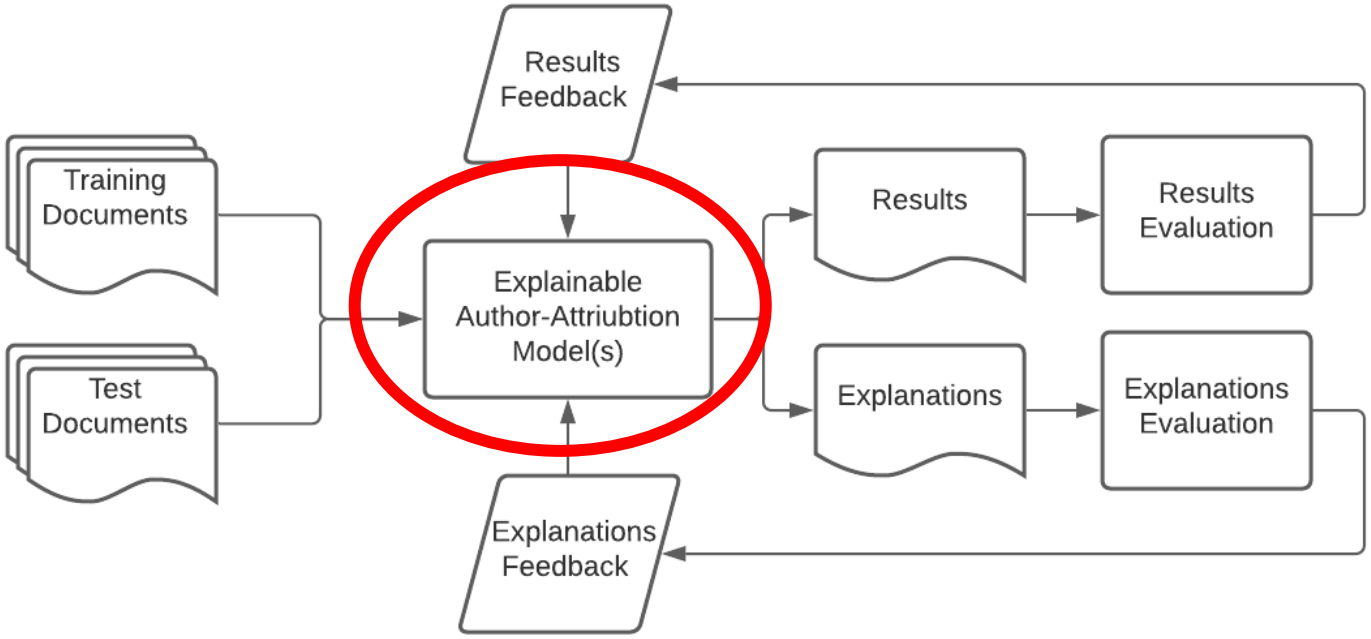
Provide high quality datasets from high priority, low-resource languages

Provide an active-learning-enhanced method for human evaluation of linguistic explanations from multiple models simultaneously

Ideal Teammate

Ideal Teammate Capabilities

- Model Building
- Explainable AI
- Software Engineering



Used by Top Brands

figure eight FEDERAL

Contact Information:

Kim Robertson
kim.robertson@f8-federal.com