Smart Crowd-Sourcing: Where Machine Learning meets Human Intelligence for Document Labelling

Antionio Penta and Gopal Ramchurn
Agents, Interaction, and Complexity Research Group
University of Southampton

Edwin Simpson and Steven Reece
Machine Learning Research Group
Oxford University

The TREC Crowdsourcing Challenge

Information Retrieval is becoming a serious challenge in the face of Big Data and the Internet of Things where information is generated by people, communities, sensors, and agents on the web.

The Crowdsourcing track of the Text Retrieval Conference has the following objectives:
-Develop crowdsourcing techniques to label 15424 documents
-Use human-machine collaboration to minimise cost and maximise efficiency
We achieved these objectives using a combination of NLP, Machine Learning, and Crowdsourcing

The challenge was to judge *18260 topic-document pairs.*

10 topics were randomly chosen by the TREC organising committee. Each topic has a *title, description and narrative* and these are used to determine if a document is relevant to a topic. Examples of topics: definition of creativity, recovery of treasure from sunken ships.

The documents come from the *TREC corpus* (TREC 8), previously labeled by NIST experts. Document sources include: Financial Times, Los Angeles Times, Federal Register *articles*

Natural Language Processing

The considered collection has around *1M distinct words*, these are extracted after a classical NLP chain made by the *lemmatization* and the *Pling-Stemming phases*. Different *strategies* are used for the features from the *word-based* count (TF-IDF, DFR) to the *cluster-based* (LDA).

A measure based on the LDA results is also used to define a relevance rank, that at the beginning helps to select the initial documents for the crowd-sourcing labeling stage.

Classification

Classifier functrion:
- identify documents to pass to the crowd
- classify documents and assign probabilities of classes

Compare *Independent Bayesian Classifier Combination* (IBCC) [1] and a traditional two-stage approach to classification

Both approaches assume independence of features extracted from text

Novelty: incorporates both learned accuracy of turkers (trust) from gold standard data and the individuals' own confidence in their responses

CrowdSourcing Techniques and Strategies

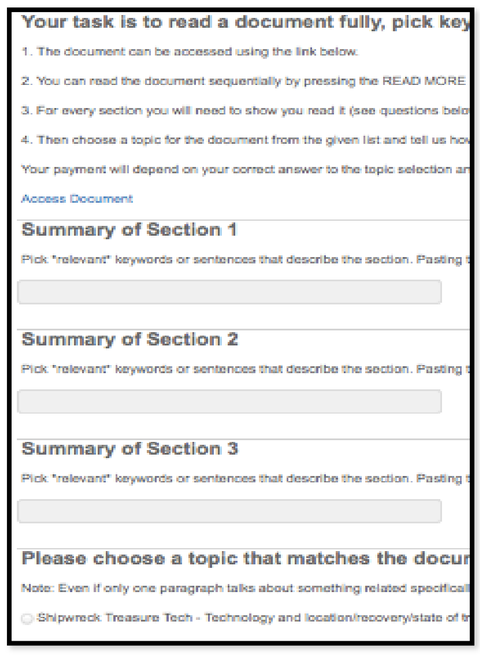
We chose to use Amazon Mechanical Turk because of the versatile and ease of use of the API that is available to this service and also because it provided access to the largest number of workers without polluting trust metrics.

To crowdsource the labelling of documents, we developed the following components:

1.HIT Interface Design –the presentation of the task to maximise efficient labelling, perform verification and avoid boredom.
2.Trust Model – functions to compute the level of trust we can assign to a turker to complete a task properly.
3.Hiring Process – workflow we used to hire workers

Trust Mechanism

- Created ‘gold’ tasks manually
- Gold task include: label + confidence in label
- Trust = (ncc+ G*ncn + (1-G)*nin+1)/(N+2)
  - ncc = no_tasks correct confident
  - ncn = no_tasks correct not confident
  - nin = no_tasks incorrect not confident
  - N = Total Tasks, G = 0.5
- Allow only those with Trust > 0.5 to work
- Disqualify those not filling in sections



Hiring Process

- Create gold tasks and compute trust
- Query Classifier for files to be labelled
- Create HITs on AMT for all chunks of documents requiring labels
- Notify workers with Trust > 0.5 that work is available
- Feed labels to classifier and get new labels

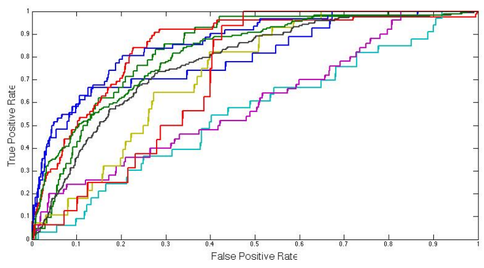
HIT Interface Design

- Split documents into small chunks (< 20KB).
- Divide each chunk into three sections to avoid information overload
- Classes redefined to be more succint and more memorable

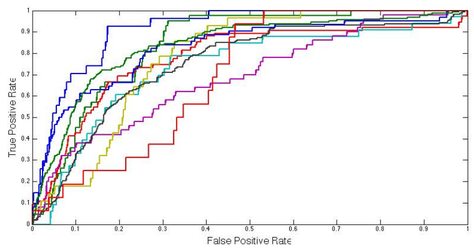
Initial Results

- > 2000 documents classified by the crowd
- 15 out of 40 turkers passed the test
- Test cost 20 dollars
- < 5 completed most tasks
- Each hit performed twice for robustness
- 1 turker blocked
- HIT cost = 0.08

Results



2 Stage (with trust)



IBCC (with confidence)

Classifier	AUC (spread)
2 Stage (without trust)	0.75 ± 0.10
2 Stage (with trust)	0.75 ± 0.11
IBCC (without confidence)	0.78 ± 0.07
IBCC (with confidence)	0.78 ± 0.07

- IBCC improves on 2 stage approach
- Confidence has no noticable impact on performance

Conclusions and References

Presented a system for document topic analysis using crowdsourcing, NLP and independent classifier combination of turker reponses.

Compared the IBCC classifier against a traditional two-stage classifier on a challenging dataset. These algorithms have been extended to incorporate turker trust and confidence measures.

Have submitted results to the TREC committee and are eagerly awaiting the results of the competition!

Please see the Orchid poster:

[1] Dynamic Bayesian Combination of Human and Artificial Decision Makers, Edwin Simpson and Stephen Roberts