Terrier Team | University of Glasgow

29TH ACM CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT
CIKM 2011
Glasgow, UK
24-28 October 2011

Crowdsourcing a News Query Classification Dataset

Richard McCreadie, Craig Macdonald & Iadh Ounis

Introduction

- **What is *news query classification* and why would we build a dataset to examine it?**
  - *Binary classification task performed by Web search engines*
  - Up to 10% of queries may be news-related [Bar-Ilan et al, 2009]
- **Have workers judge Web search *queries* as *news-related* or not**

Introduction

But:

- **News-relatedness is subjective**
- **Workers can easily `game` the task**
- **News Queries change over time**

```
for  query in task
                return Random(Yes,No)
end loop
```

How can we overcome these difficulties to create a high quality dataset for news query classification?

Talk Outline

Terrier Team | University of Glasgow

- ✓ Introduction (1--3)
- ❑ Dataset Construction Methodology (4--14)
- ❑ Research Questions and Setting (15--17)
- ❑ Experiments and Results (18--20)
- ❑ Conclusions (21)

Dataset Construction Methodology

- **How can we go about building a news query classification dataset?**

1. Sample Queries from the MSN May 2006 Query Log
2. Create Gold judgments to validate the workers
3. Propose additional content to tackle the temporal nature of news queries and prototype interfaces to evaluate this content on a small test set
4. Create the final labels using the best setting and interface
5. Evaluate in terms of agreement
6. Evaluate against `experts`

Dataset Construction Methodology

- **Sampling Queries:**
  - Create **2** query sets sampled from the MSN May 2006 Query-log
    - Poisson Sampling
  - One for testing (***testset***)
    - Fast crowdsourcing turn around time
    - Very low cost
  - One for the final dataset (***fullset***)
    - 10x queries
    - Only labelled once

Date	Time	Query
2006-05-01	00:00:08	What is May Day?
2006-05-08	14:43:42	protest in Puerto rico

Testset Queries : 91

Fullset Queries : 1206

Dataset Construction Methodology

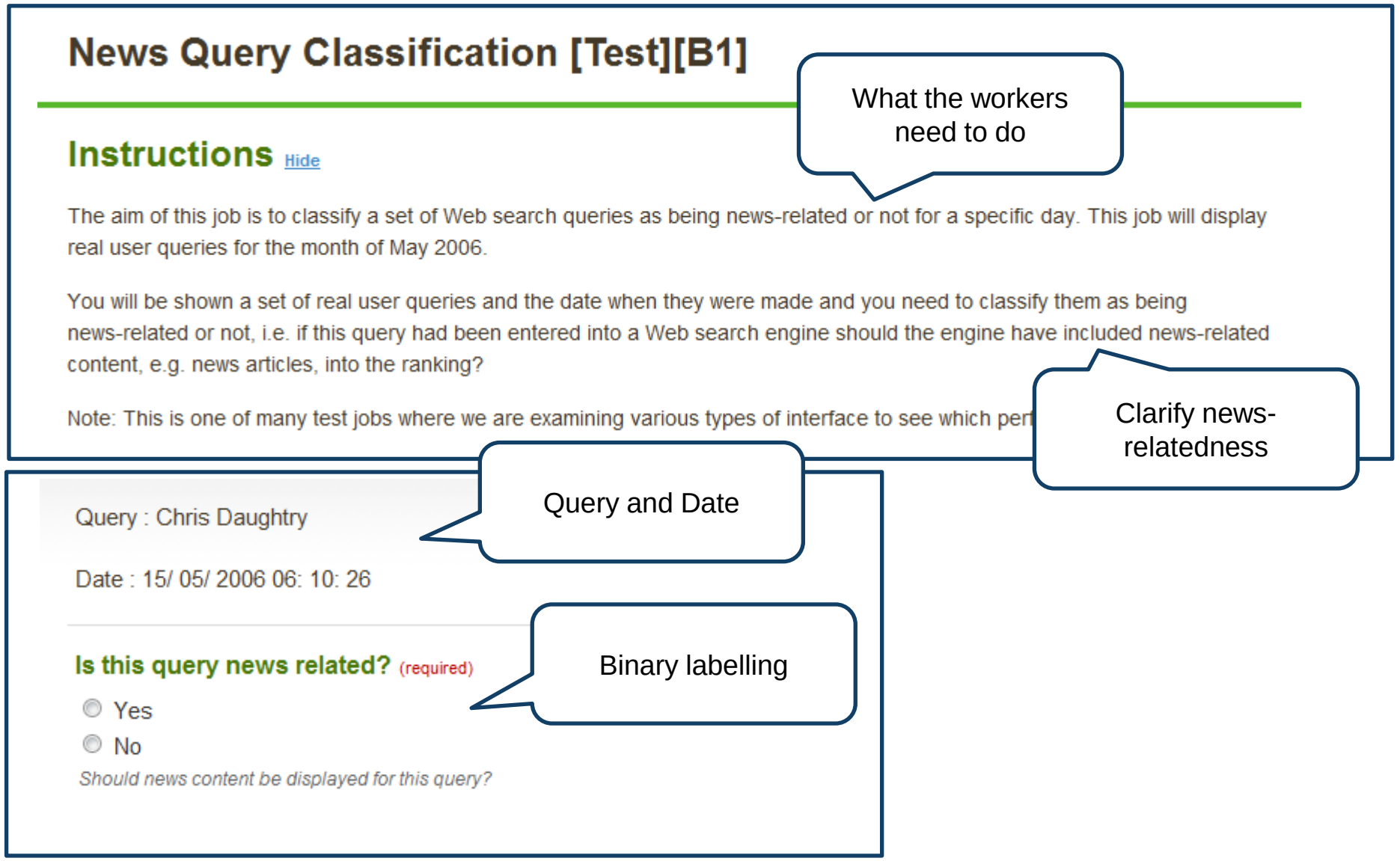
- **How to check our workers are not `gaming' the system?**
  - **Gold Judgments (honey-pot)**
    - Small set (5%) of queries
    - Catch out bad workers early in the task
    - `Cherry-picked` unambiguous queries
    - Focus on news-related queries
  - **Multiple workers per query**
    - 3 workers
    - Majority result

Date	Time	Query	Validation
2006-05-01	00:00:08	What is May Day?	No
2006-05-08	14:43:42	protest in Puerto rico	Yes

Dataset Construction Methodology

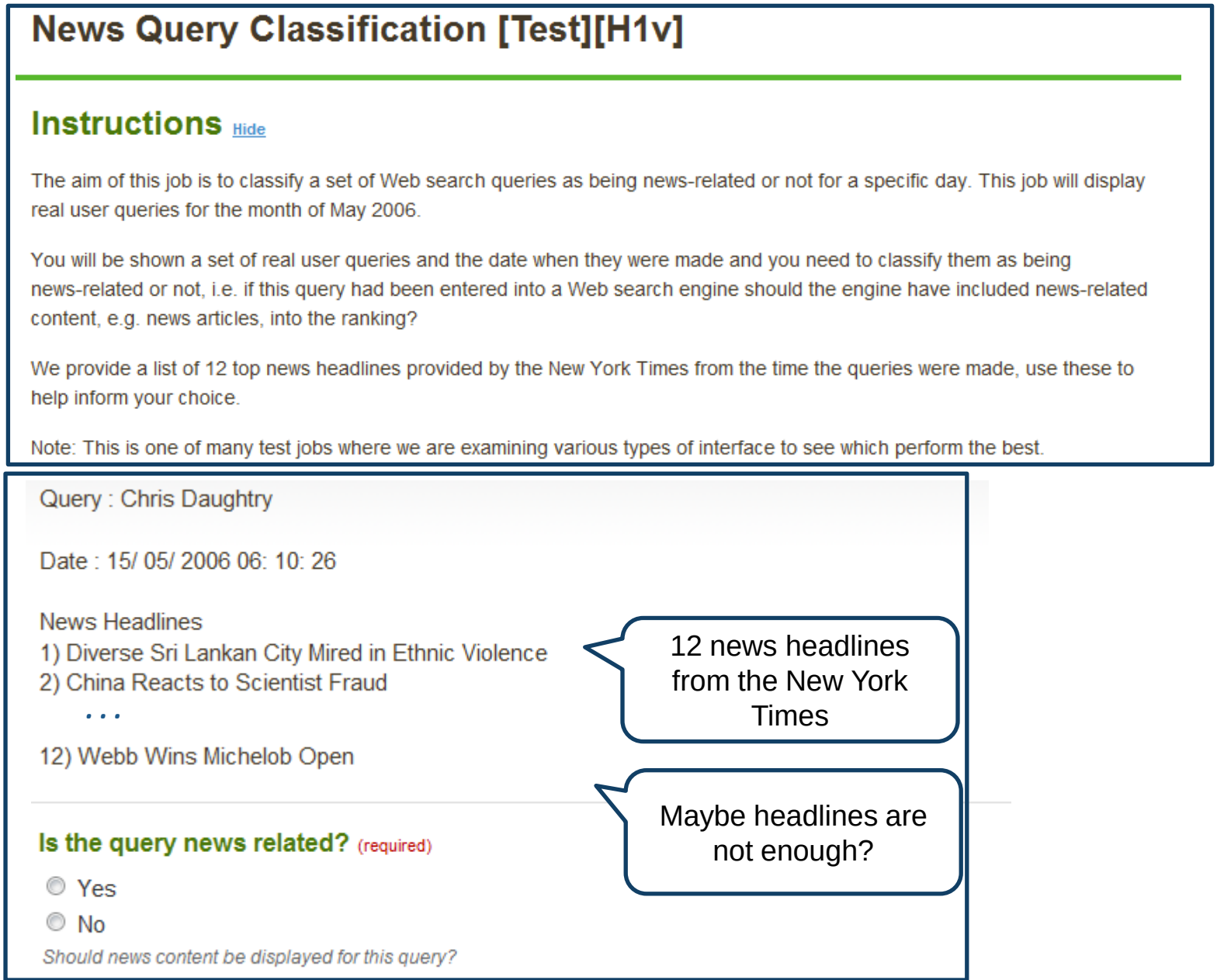
- **How to counter temporal nature of news queries?**
  - Workers need to know what the news stories of the time were . . .
  - But likely will not remember what the main stories during May 2006
- **Idea: add extra information to interface**
  - News Headlines
  - News Summaries
  - Web Search results
- **Prototype Interfaces**
  - Use small testset to keep costs and turn-around time low
  - See which works the best

Interfaces : Basic

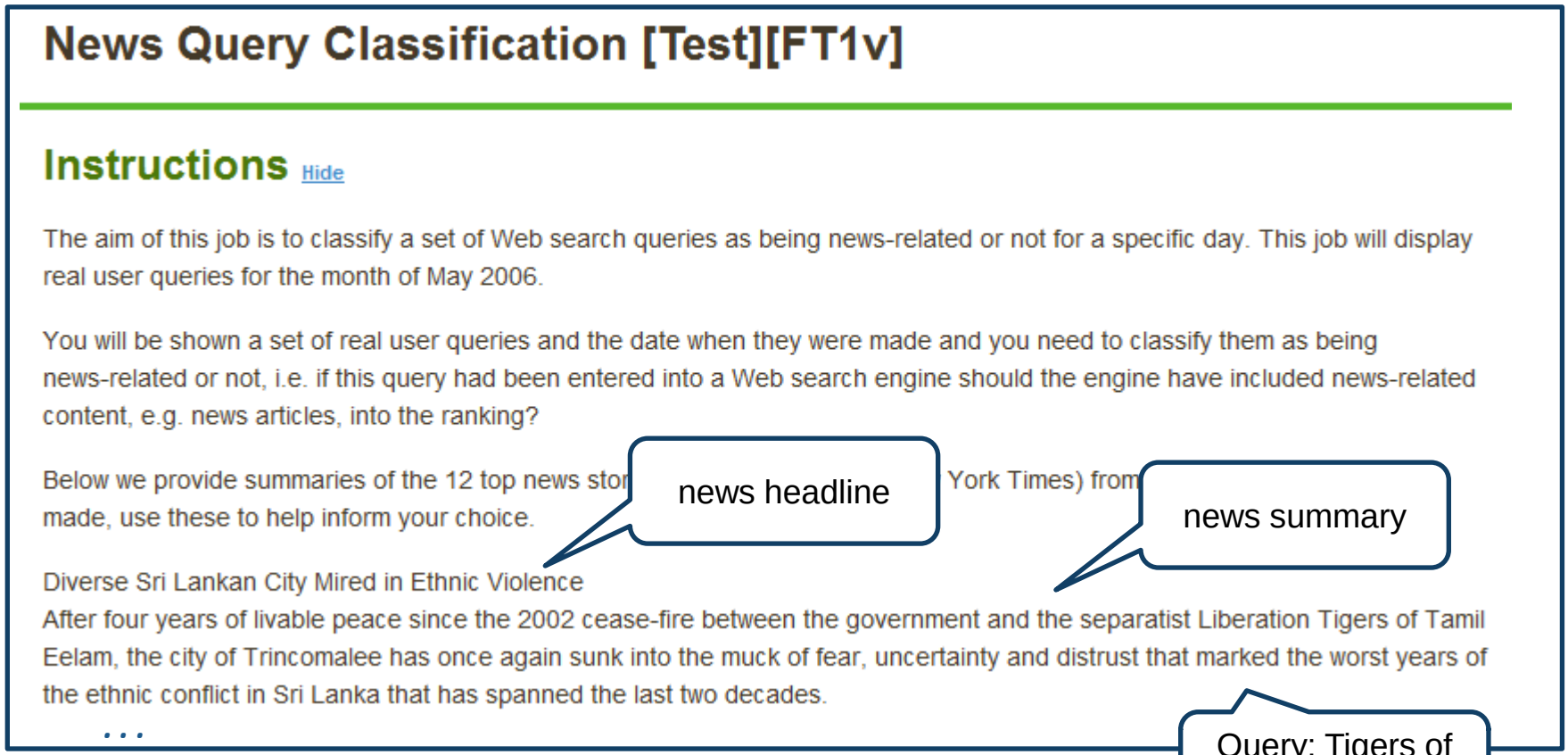


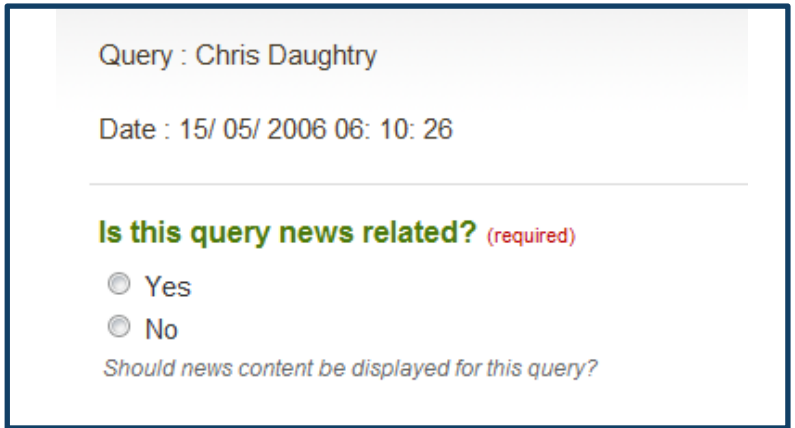
Interfaces : Headline

Interfaces : HeadlineInline



Interfaces : HeadlineSummary





Interfaces : LinkSupported

News Query Classification [Test][L1v]

Instructions Hide

The aim of this job is to classify a set of Web search queries as being news-related or not for a specific day. This job will display real user queries for the month of May 2006.

You will be provided with automaticly generated links to three major search engines for each query. Browse the search results from these three search engines and then indicate whether the query is news-related or not. If you think it is news-related, then also provide a link to a web page supporting this decision from the search results displayed.

Note: This is one of many test jobs where we are examining various types of interface to see which perform the best.

Query : Chris Daughtry

Date : 15/ 05/ 2006 06: 10: 26

Link 1 : Click here

Link 2 : Click here

Link 3 : Click here

Links to three major search engines

Is the query news related? (required)

- Yes
- No

Should news content be displayed for this query?

Triggers a search containing the query and its date

If Yes, supply a link to a search result that supports your decision

for example a news article from a known newswire provider like CNN, BBC, NYTimes etc.

Also get some additional feedback from workers

Dataset Construction Methodology

- **How do we evaluate our the quality of our labels?**
  - **Agreement** between the three workers per query
    - The more the workers agree, the more confident that we can be that our resulting majority label is correct
  - Compare with `**expert**' (me) judgments
    - See how many of the queries that the workers judged news-related match the ground truth

Date	Time	Query	Worker	Expert
2006-05-05	07:31:23	abcnews	Yes	No
2006-05-08	14:43:42	protest in Puerto rico	Yes	Yes

Talk Outline

Terrier Team | University of Glasgow

Experimental Setup

- **Research Questions**
  - How do our interface and setting effect the quality of our labels
    1. Baseline quality? How bad is it?
    2. How much can the honey-pot bring?
    3. How about our extra information {headlines,summaries,result rankings}

testset

  - Can we create a good quality dataset**?**
    1. Agreement?
    2. Vs ground truth

fullset

Experimental Setup

- **Crowdsourcing Marketplace**
  - amazon mechanical turk Artificial Artificial Intelligence
- **Portal**
  - CrowdFlower
- **Judgments per query : 3**
- **Costs (per interface)**
  - Basic $1.30
  - Headline $4.59
  - HeadlineInline $4.59
  - HeadlineSummary $5.56
  - LinkSupported $8.78

- **Restrictions**
  - USA Workers Only
  - 70% gold judgment cutoff
- **Measures**
  - Compare with ground truth
    - **Precision, Recall, Accuracy** over our expert ground truth
  - Worker agreement
    - Free-marginal multirater Kappa (***Kfree***)
    - Fleiss multirater Kappa (***Kfleiss***)

Talk Outline

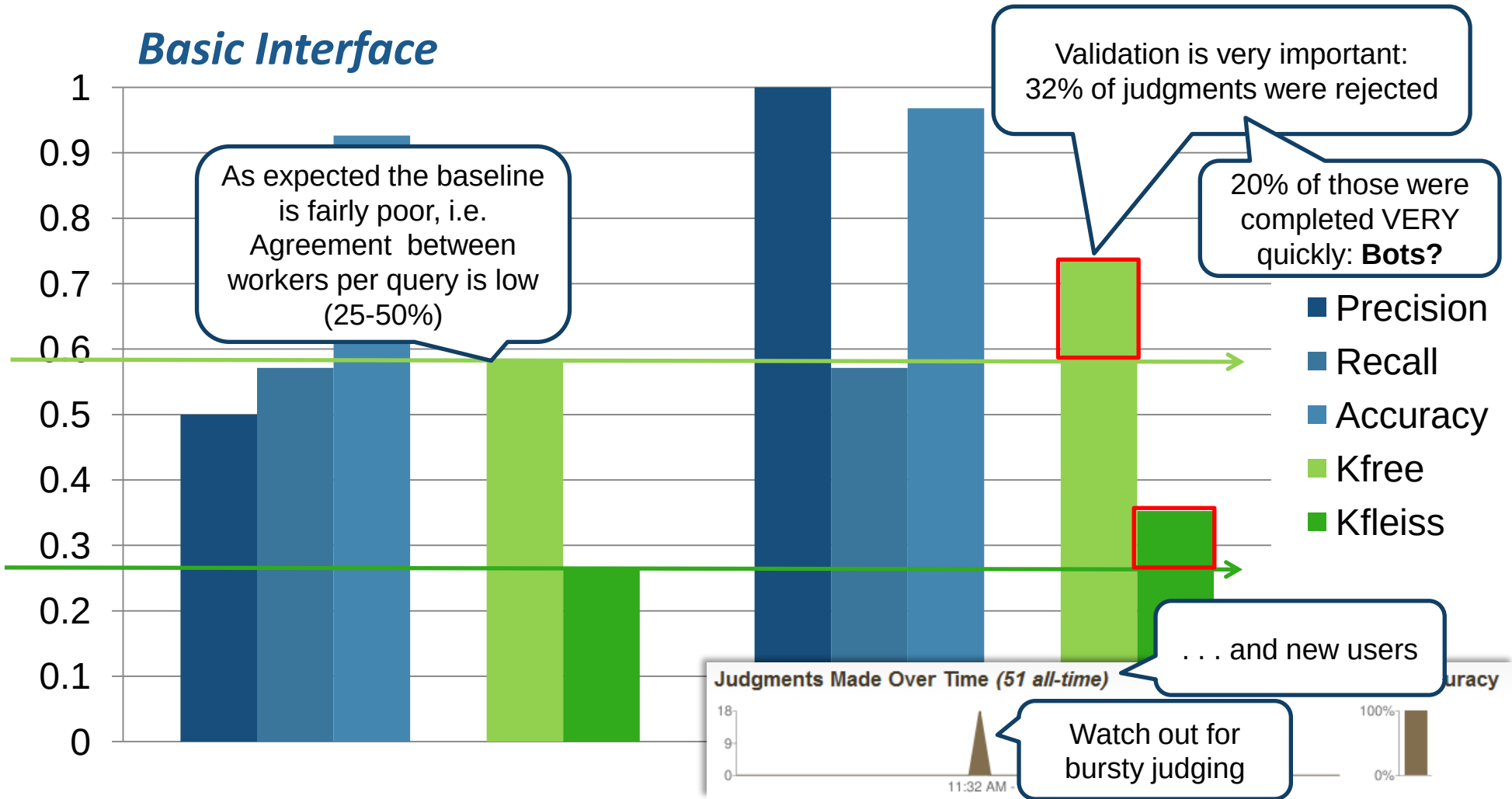
Terrier Team | University of Glasgow

- ☑ Introduction (1--3)
- ☑ Dataset Construction Methodology (4--14)
- ☑ Research Questions and Setting (15--17)
- ☐ Experiments and Results (18--20)
- ☐ Conclusions (21)

Baseline and Validation

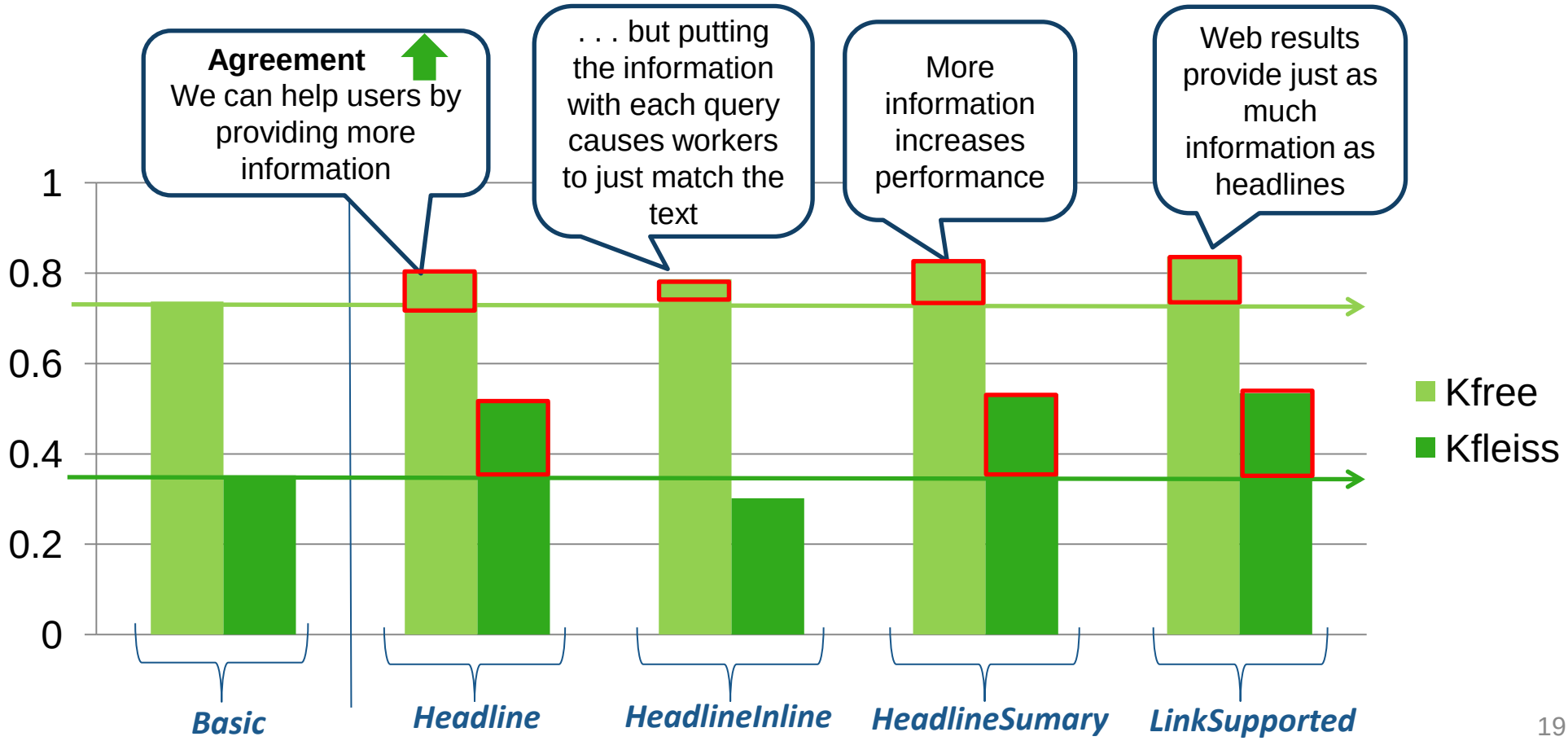
Terrier Team | University of Glasgow

- How is our Baseline?
- What is the effect of validation?



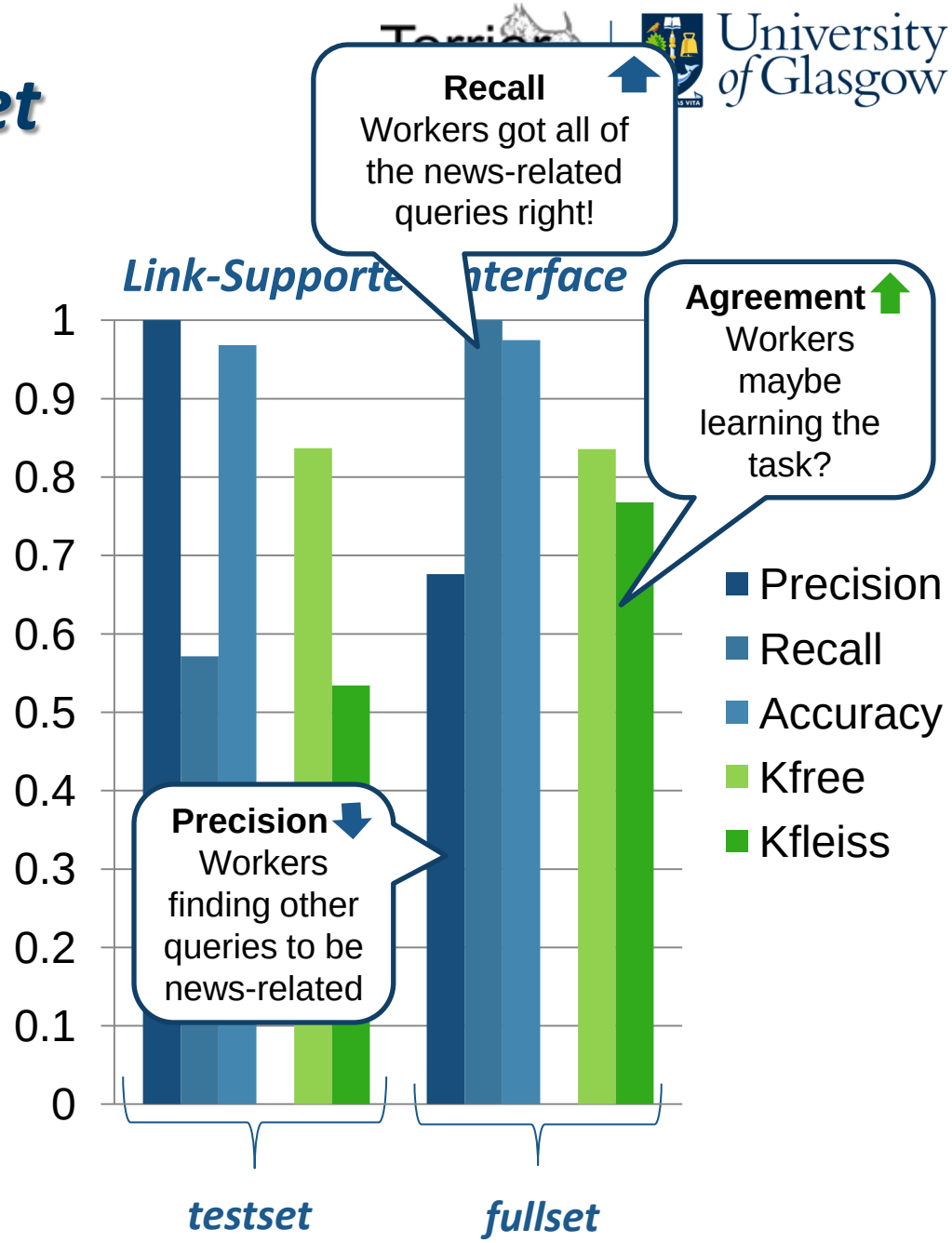
Adding Additional Information

- By providing additional news-related information does label quality increase and which is the best interface



Labelling the FullSet

- We now label the fullset
  - 1204 queries
  - Gold Judgments
  - LinkSupported Interface
- Are the resulting labels of sufficient quality?
- High recall and agreement indicate that the labels are of high quality



Conclusions & Best Practices

- **Crowdsourcing is useful for building a news-query classification dataset**
- **We are confident that our dataset is reliable since agreement is high**
  - Best Practices
    - Online worker validation is paramount, catch out bots and lazy workers to improve agreement
    - Provide workers with additional information to help improve labelling quality
    - Workers can learn, running large single jobs may allow workers to become better at the task

Questions?