

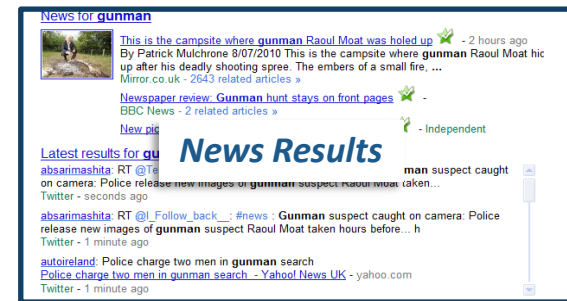
# Crowdsourcing a News Query Classification Dataset

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
```

News-related?

Query:  → Sea Creature?

Octopus?  → Work Cup Predications?

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

# Talk Outline

- 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

## News Query Classification [Test][B1]

### 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 shown a set of real user queries and the date when they were made and you need to classify them as being news-related or not, i.e. if this query had been entered into a Web search engine should the engine have included news-related content, e.g. news articles, into the ranking?

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

What the workers need to do

Clarify news-relatedness

Query : Chris Daughtry

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

**Is this query news related?** (required)

Yes

No

*Should news content be displayed for this query?*

Query and Date

Binary labelling

# Interfaces : *Headline*

## News Query Classification [Test][H1v]

### 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 shown a set of real user queries and the date when they were made and you need to classify them as being news-related or not, i.e. if this query had been entered into a Web search engine should the engine have included news-related content, e.g. news articles, into the ranking?

Below we provide a list of 12 top news headlines provided by the New York Times from the time the queries were made, use these to help inform your choice.

News Headlines:

- 1) Diverse Sri Lankan City Mired in Ethnic Violence
- 2) China Reacts to Scientist Fraud

...

12 news headlines  
from the New York  
Times

Will the workers  
bother to read these?

Query : Chris Daughtry

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

**Is this query news related?** *(required)*

- Yes  
 No

*Should news content be displayed for this query?*

# Interfaces : HeadlineInline

## News Query Classification [Test][H1v]

### 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 shown a set of real user queries and the date when they were made and you need to classify them as being news-related or not, i.e. if this query had been entered into a Web search engine should the engine have included news-related content, e.g. news articles, into the ranking?

We provide a list of 12 top news headlines provided by the New York Times from the time the queries were made, use these to help inform your choice.

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

News Headlines

- 1) Diverse Sri Lankan City Mired in Ethnic Violence
- 2) China Reacts to Scientist Fraud

...

12) Webb Wins Michelob Open

12 news headlines  
from the New York  
Times

Maybe headlines are  
not enough?

**Is the query news related?** (required)

- Yes
- No

Should news content be displayed for this query?

# Interfaces : *HeadlineSummary*

## News Query Classification [Test][FT1v]

### 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 shown a set of real user queries and the date when they were made and you need to classify them as being news-related or not, i.e. if this query had been entered into a Web search engine should the engine have included news-related content, e.g. news articles, into the ranking?

Below we provide summaries of the 12 top news stories made, use these to help inform your choice.

news headline

(York Times) from

news summary

Diverse Sri Lankan City Mired in Ethnic Violence

After four years of livable peace since the 2002 cease-fire between the government and the separatist Liberation Tigers of Tamil Eelam, the city of Trincomalee has once again sunk into the muck of fear, uncertainty and distrust that marked the worst years of the ethnic conflict in Sri Lanka that has spanned the last two decades.

...

Query: Tigers of  
Tamil?

Query : Chris Daughtry

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

**Is this query news related?** (required)

- Yes
- No

*Should news content be displayed for this query?*

# 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 automatically 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](#)

Is the query news related? (required)

Yes

No

Should news content be displayed for this query?

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.

Links to three major search engines

Triggers a search containing the query and its date

The screenshot shows a Google search interface. On the left, there are filters for 'Everything', 'Videos', and 'More'. Below these is a 'Show search tools' link. The main search results area shows the search query '15 May 2006 Chris Daughtry' with 'About 92,700 results (0.24 seconds)'. The first result is 'Chris Daughtry' with a star icon and a 'Verified' badge. The snippet reads: 'American Idol: Chris Daughtry -- Chris Daughtry didn't need a victory on .... Episode dated 15 May 2006 (2006) TV episode .... Himself, "What Perez Sez" . ... www.imdb.com/name/nm0202322/ - Cached - Similar'. The second result is '"Live with Regis" Episode dated 15 May 2006 (2006)' with a star icon and a 'Verified' badge. The snippet reads: 'With Chris Daughtry, Jeff Probst, Terry Quinn. Visit IMDb for Photos, Showtimes, ... with Regis and Kathie Lee" Episode dated 15 May 2006 (original title) ... www.imdb.com/title/tt0803868/ - Cached'. The third result is 'Chris Daughtry Icontest' with a star icon. The snippet reads: '[06] Voting! (15 May 2006|05:34pm) ... The theme for this week is: Chris Outside of American Idol. .... music, |, Daughtry || Emotion, | ... community.livejournal.com/daughtrycontest - Cached - Similar'.

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

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

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

*testset*

*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

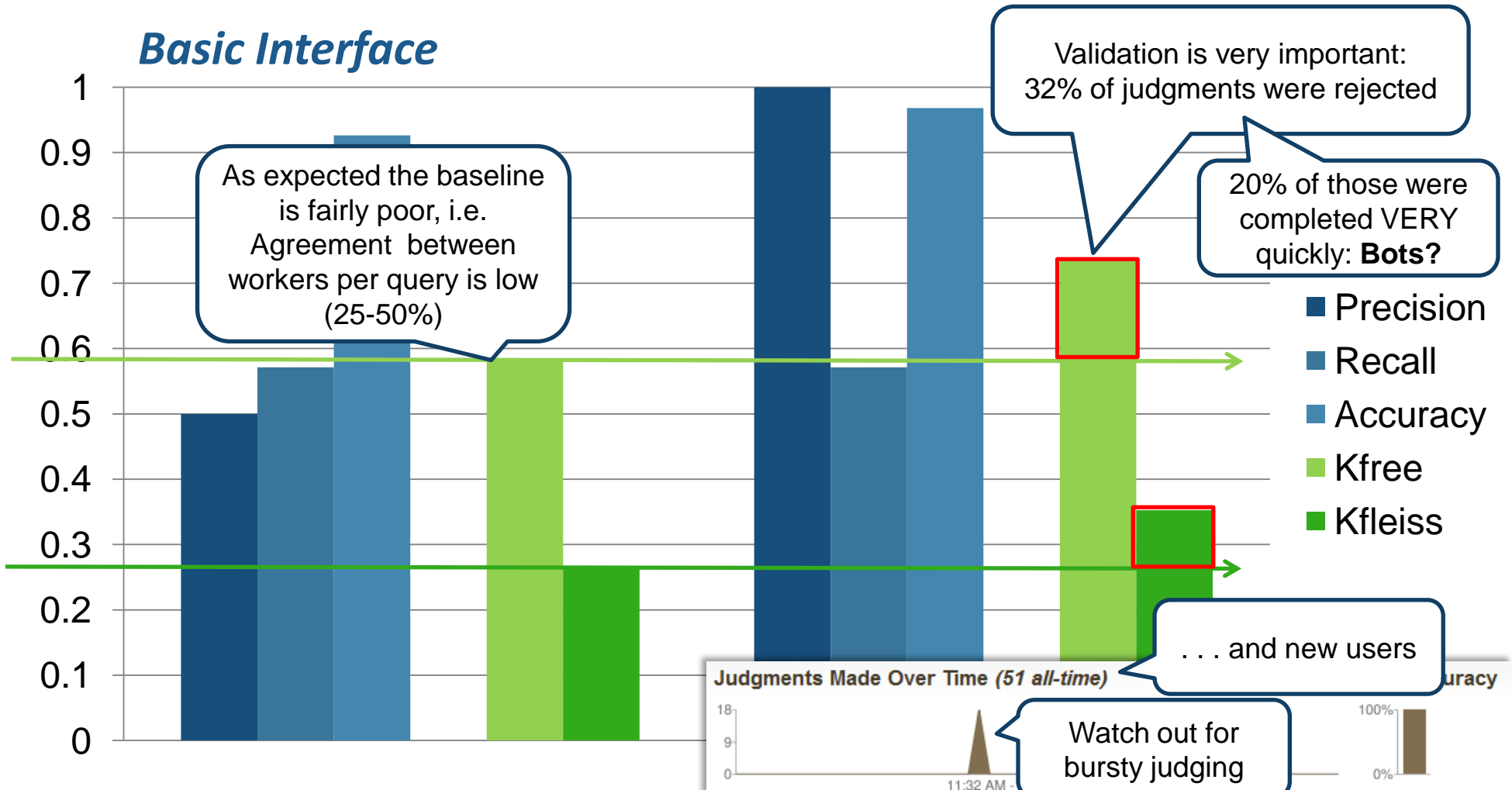
- Introduction (1--3)
- Dataset Construction Methodology (4--14)
- Research Questions and Setting (15--17)
- Experiments and Results (18--20)
- Conclusions (21)

# Baseline and Validation

## • How is our Baseline?

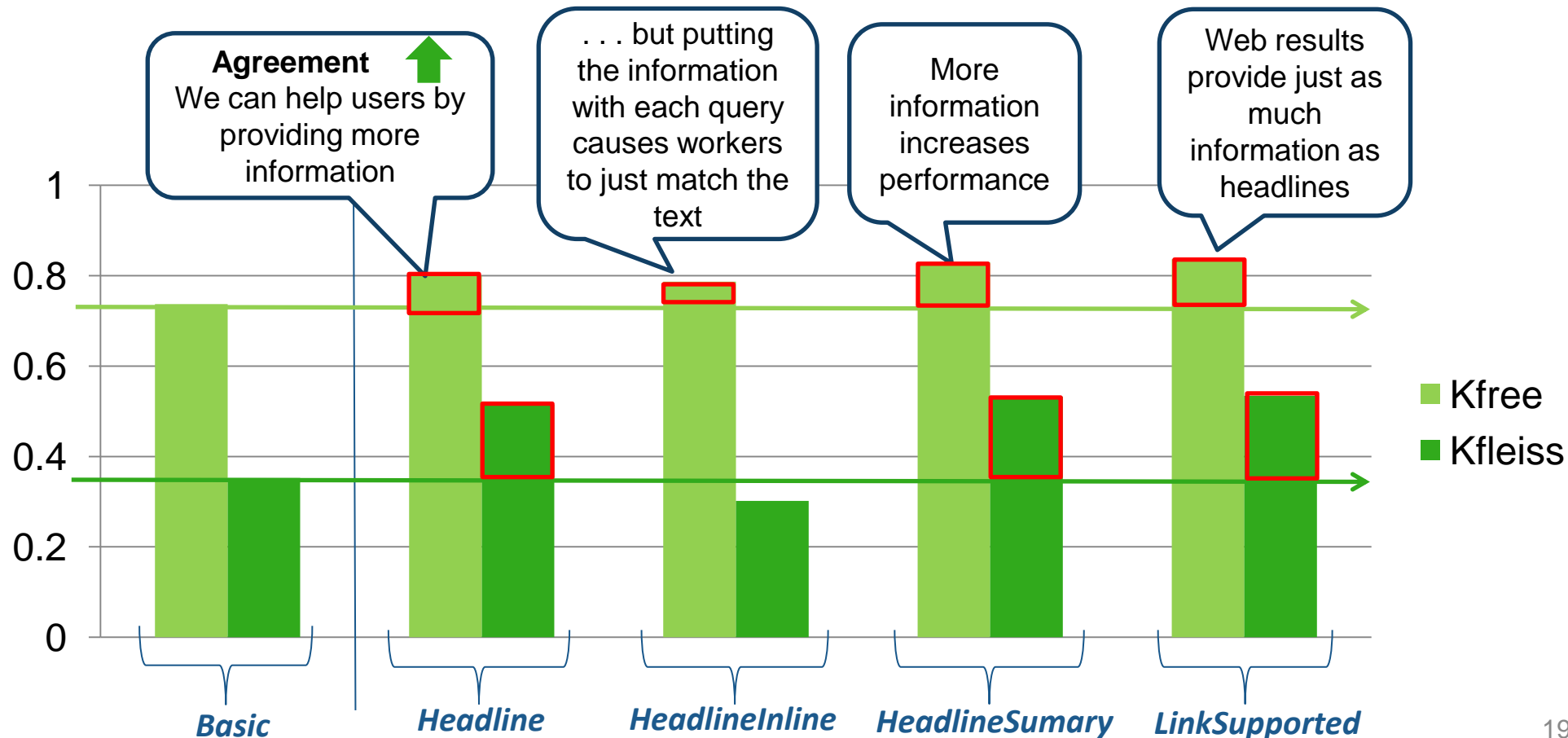
## • What is the effect of validation?

### Basic Interface



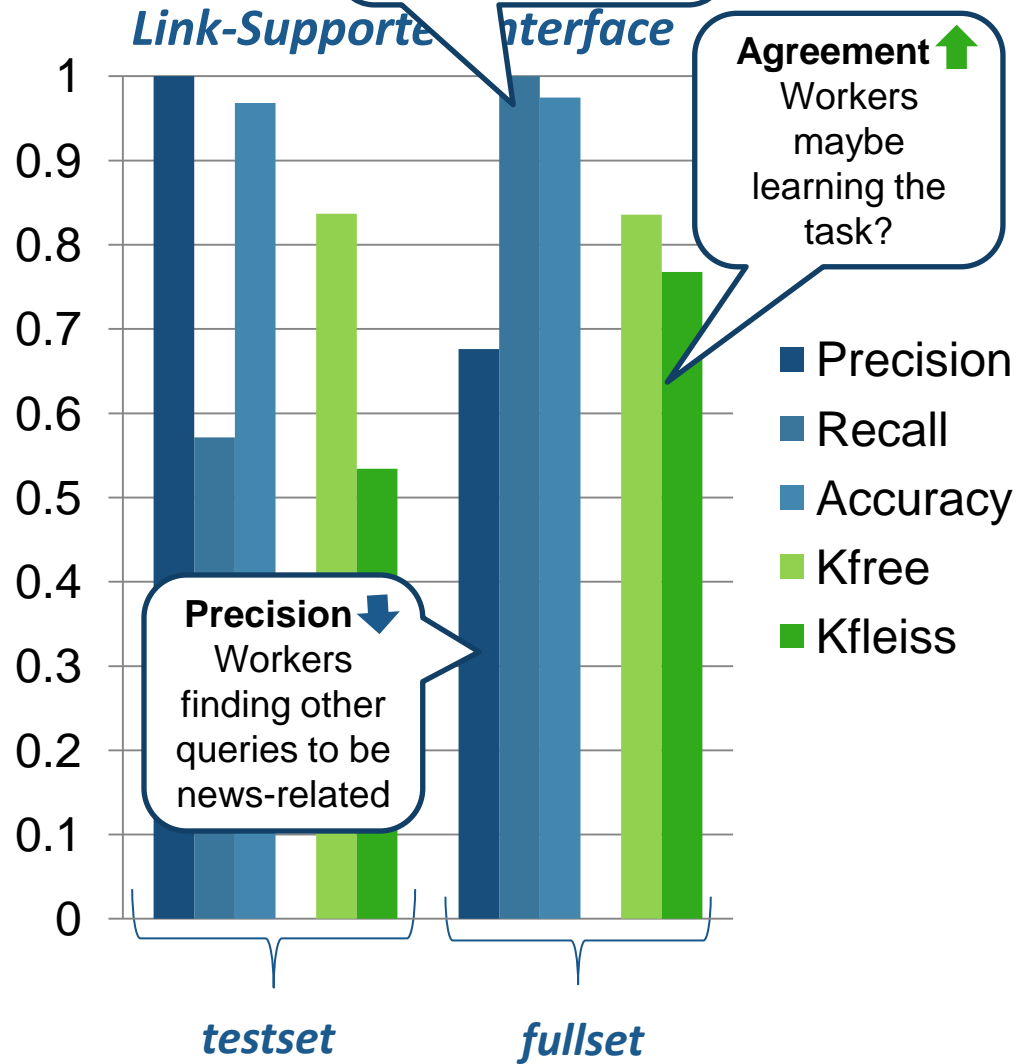
# Adding Additional Information

- 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?***