University of Sheffield,
NLP

Crowdsourcing Social Media Corpora

Kalina Bontcheva
Leon Derczynski

© The University of Sheffield, 1995-2014
This work is licensed under
the Creative Commons Attribution-NonCommercial-NoDerivs Licence

Why Annotate New Social Media Corpora?

- Plenty of already annotated corpora in the news and similar genres
- Big enough for both training and evaluation
- Social media corpora annotated for many NLP tasks are unfortunately largely lacking or too small in comparison to their news counterparts
- We will look into how best to create these in an affordable manner

- LREC 2014 paper: "Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines" Sabou, Bontcheva, Derczynski, Scharl

The Science of Corpus Annotation

- Quite well understood best practice in how to create linguistic annotation of consistently high quality by employing, training, and managing groups of linguistic and/or domain experts
- Necessary in order to ensure reusability and repeatability of results
- The acquired corpora are of very high quality
- Costs are unfortunately also very high: estimated at between $0.36 and $1.0 (Zaidan and Callison-Burch, 2011; Poesio et al., 2012)

What is Crowdsourcing?

- Crowdsourcing is an emerging collaborative approach for acquiring annotated corpora and a wide range of other linguistic resources
- Three main kinds of crowdsourcing platforms
  - paid-for marketplaces such as Amazon Mechanical Turk (AMT) and CrowdFlower (CF)
  - games with a purpose
  - volunteer-based platforms such as crowdcrafting

Why Crowdsourcing?

- Paid for crowdsourcing can be 33% cheaper than in-house employees when applied to tasks such as tagging and classification (Hoffmann, 2009)
- Games with a purpose can be even cheaper in the long run, since the players are not paid.
- However cost of implementing a game can be higher than AMT/CF costs for smaller projects (Poesio et al, 2012)
- Tap into the large number of contributors/players available across the globe, through the internet
- Easy to reach native speakers in various languages (but beware Google translate cheaters!)

Genre 1: Mechanised Labour

- Participants (workers) paid a small amount of money to complete easy tasks (HIT = Human Intelligence Task)

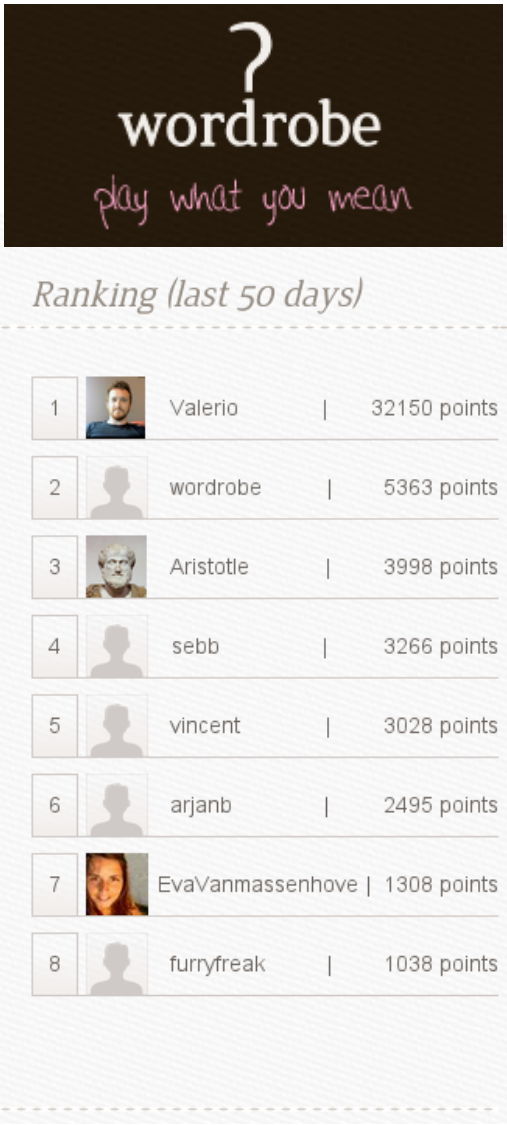
amazon mechanical turk™
Artificial Artificial Intelligence

CrowdFlower

Paid for Crowdsourcing

- Contributors are extrinsically motivated through economic incentives
- Carry out micro-tasks in return for micro-payments
- Most NLP projects use crowdsourcing marketplaces: Amazon Mechanical Tutk and CrowdFlower
- Requesters post Human Intelligence Tasks (HITs) to a large population of micro-workers (Callison-Burch and Dredze, 2010a)
- Snow et al. (2008) collect event and affect annotations, while Lawson et al. (2010) and Finin et al. (2010) annotate special types of texts such as emails and Twitter feeds, respectively.
- Challenges:
  - low quality output due to the workers' purely economic motivation
  - high costs for large tasks (Parent and Eskenazi, 2011)
  - ethical issues (Fort et al., 2011)

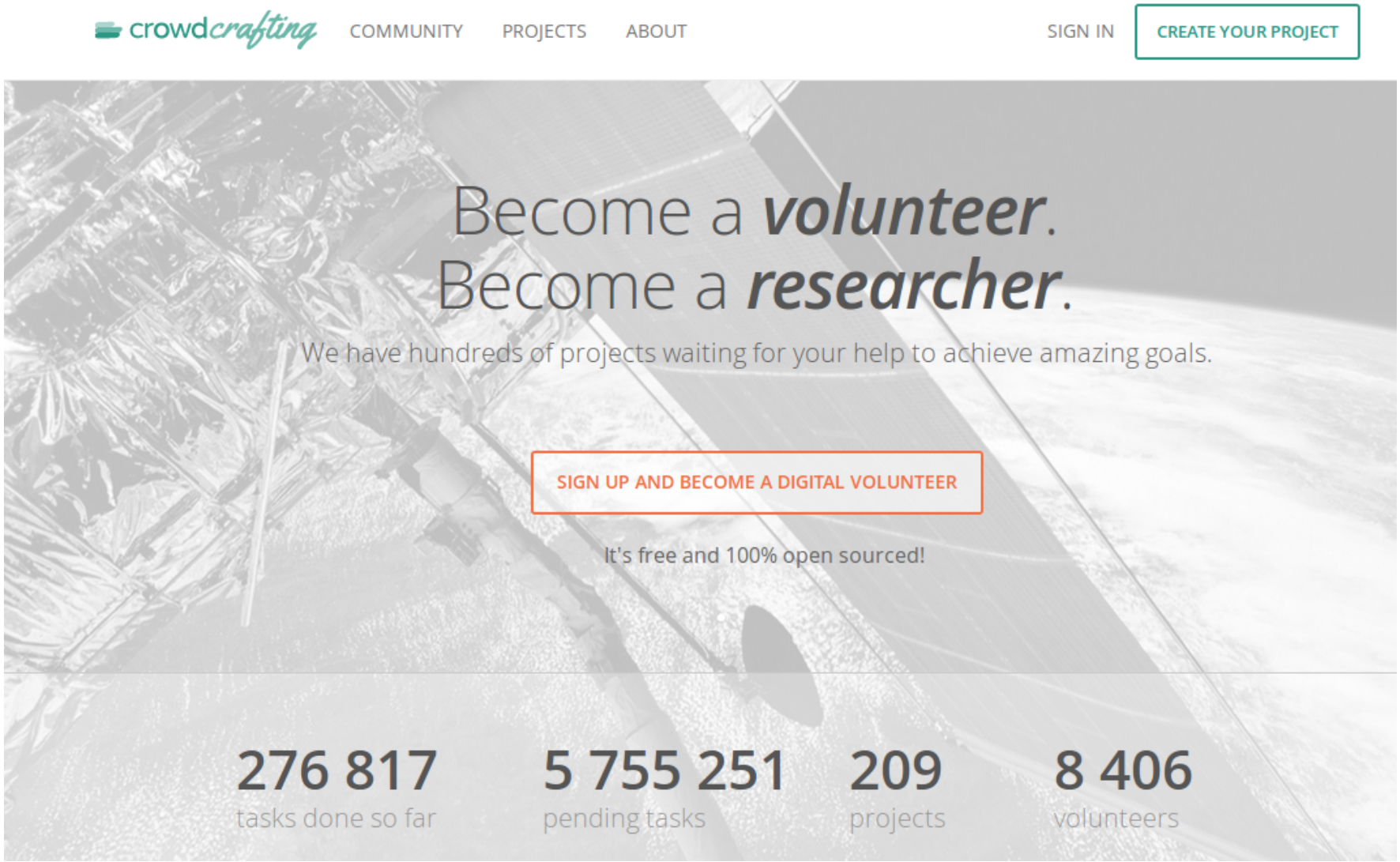
Genre 2: Games with a purpose (GWAPs)



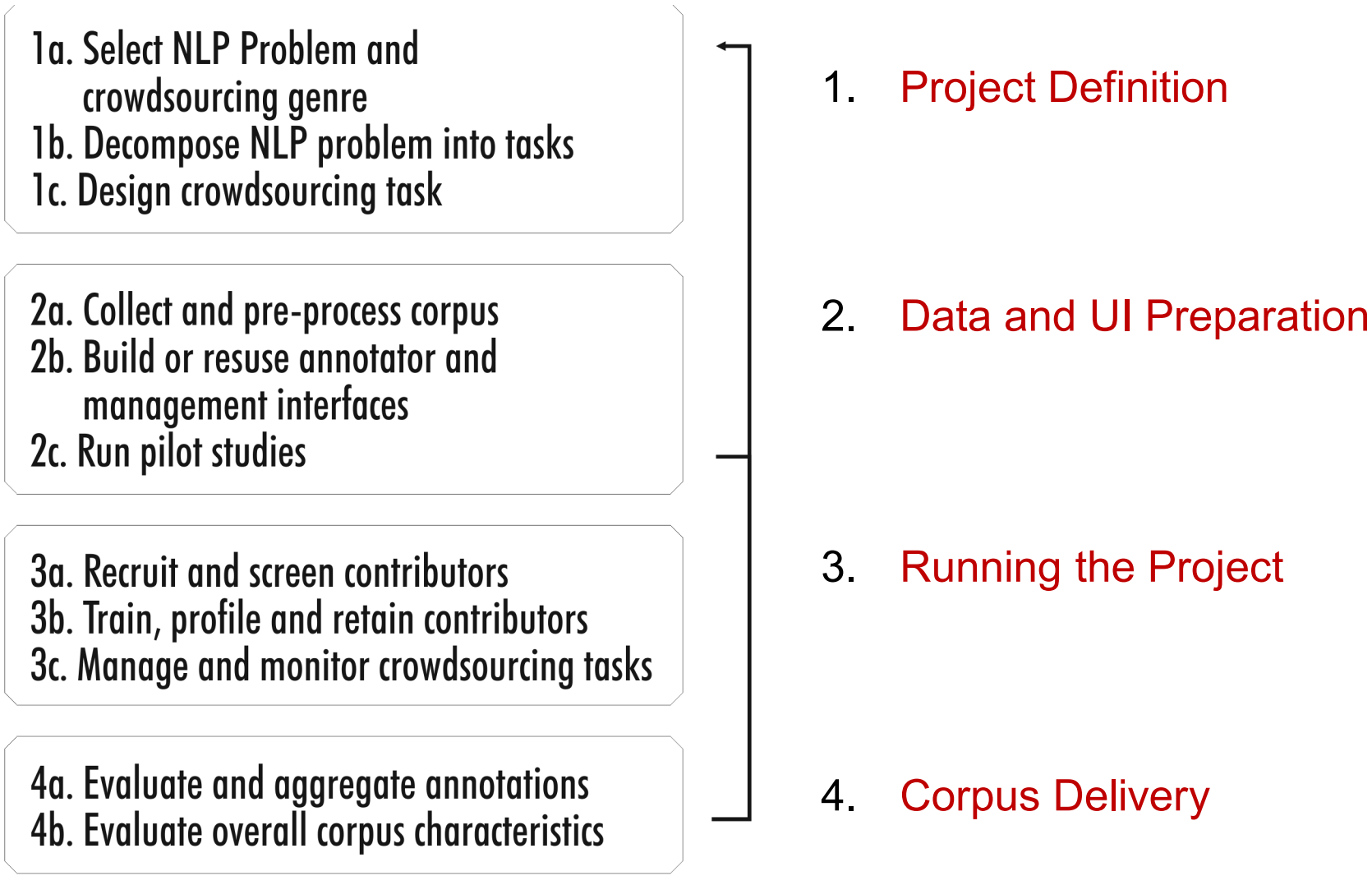
Games with a Purpose (GWAPs)

- In GWAPs (von Ahn and Dabbish, 2008), contributors carry out annotation tasks as a side effect of playing a game
- Compared to paid-for marketplaces, GWAPs:
  - reduce costs and the incentive to cheat as players are intrinsically motivated
  - promise superior results, due to motivated players and better utilization of sporadic, explorer-type users (Parent and Eskenazi, 2011)
- Example GWAPs:
  - Phratris for annotating syntactic dependencies (Attardi, 2010)
  - PhraseDetectives (Poesio et al.,2012) to acquire anaphora annotations
  - Sentiment Quiz (Scharl et al., 2012) to annotate sentiment
  - http://www.wordrobe.org/ - A collection of NLP games incl. POS, NE
- Challenges:
  - Designing pealing games and attracting a critical mass of players are among the key success factors within this genre (Wang et al., 2012)

Genre 3: Altruistic Crowdsourcing



Workflow for Crowdsourcing Corpora



Step 1. Project Definition

- Data distribution: how “micro” is each microtask?
  - Long paragraphs hard to digest, worker fatigue
  - Single sentences not always appropriate: e.g. for co-ref
- Reward scheme
  - Granularity – per task? Per set of tasks? High scores?
  - What to do with “bad” work
  - How much to reward
    - No clear, repeatable results for quality:reward relation
    - High rewards get it done faster, but not better
    - Pilot task gives timings, so pay at least minimum wage

Step 1. Project Definition

- Choose the most appropriate genre or mixture of crowdsourcing genres
  - Trade-offs: Cost; Timescale; Worker skills

- Pilot the the design, measure performance, try again
  - Simple, clear design important
  - Binary decision tasks get good results

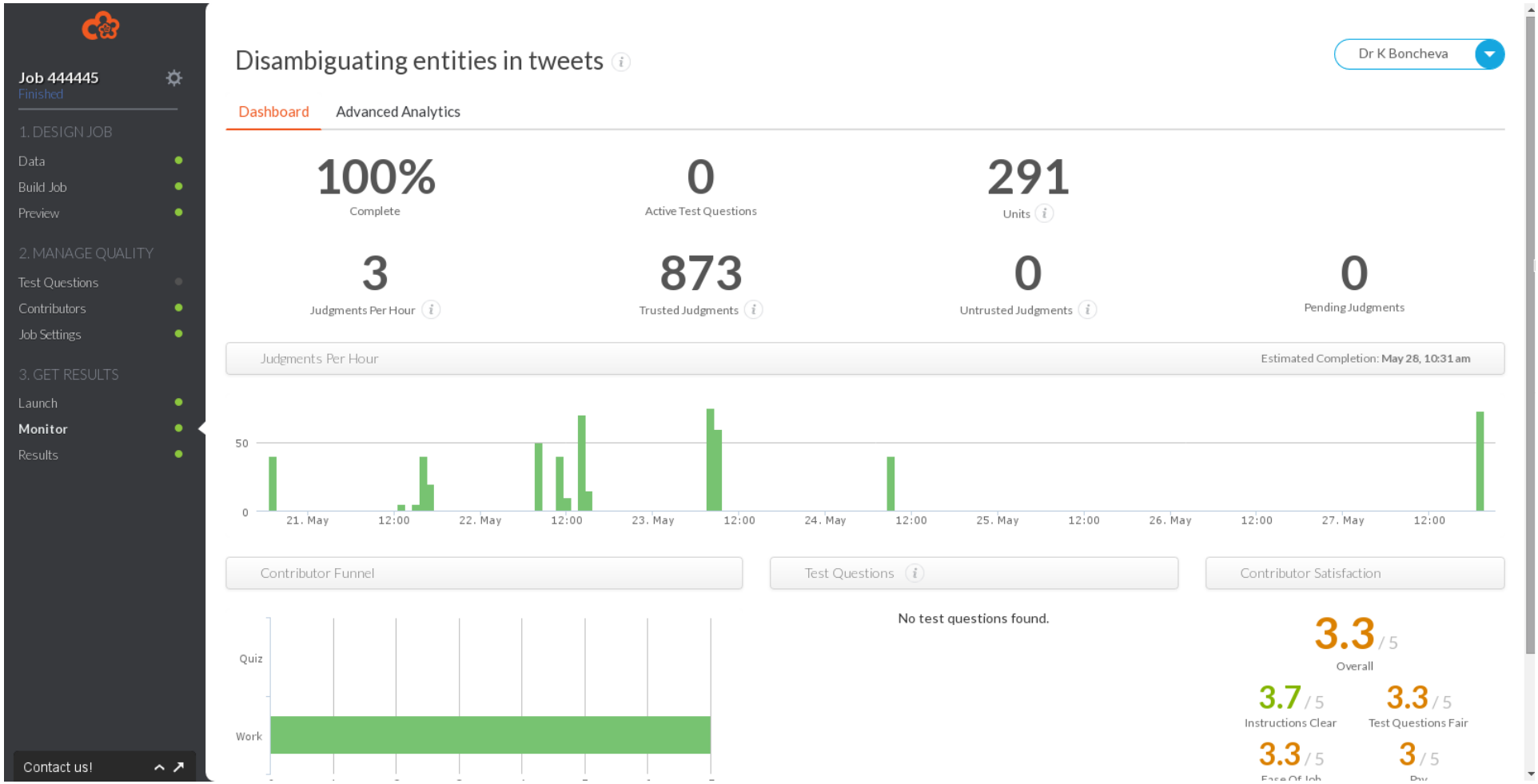
Step 1. Project Definition

- Named entity recognition example:
  - Entity selection options
    - Allow users to select entities with the mouse
    - Ask users to click on the words which constitute the entity
    - Show users a highlighted entity in context and ask them to classify its type
  - Task definition options
    - Ask users to classify entities into 4-7 pre-defined classes simultaneously
    - Focus on one entity class only, e.g. locations, and ask users to mark only these
  - Distinguish tweets with no entities from tweets where user has not marked anything

Step 2: Data and UI Preparation

- Pre-process the corpus linguistically, as needed, e.g.
  - Tokenise text if user needs to select words
  - Identify proper names/noun phrases if we want to classify these
  - Bring additional context, if needed, e.g. text of user profile from Twitter
- Build and test the user interfaces
  - Easy to medium difficulty in AMT/CF and crowdcrafting
  - Medium to hard for GWAPs
- Run bigger pilot studies with volunteers to test everything and collect gold units for quality control later

Step 2: Data and UI Preparation



Step 3: Running the Crowdsourcing Project

- Can run for hours, days or years, depending on genre and size
- Task workflow and management
  - Create/verify workflows where challenging NLP tasks are decomposed into simpler ones. Where disagreement exists, the task is sent to be verified by another set of annotators
  - E.g., if “Manchester” is marked as a location by some contributors and as referring to an organisation by others (e.g. Manchester United FC), then show the 2 alternatives to new contributors asking them which is correct in the given context
- Contributor management (including profiling and retention)
  - Recruit volunteers (e.g. restrict by country/spoken language, advertise in media)
  - Test their knowledge, if needed
  - Have sufficient number of contributors
    - Lawson et al. (2010): number of required labels varies for different aspects of the same NLP problem. Good results with only four annotators for Person NEs, but require six for Location and seven for Organizations
- Quality control
  - Use gold units to control quality

Step 3: Running the Crowdsourcing Project

- Multi-batch methodology
  - Submit tasks in multiple batches
  - Restrictions by country/language
  - Contributor diversity by starting batches at different times
  - Needs less gold data

Step 4: Evaluation and Corpus Delivery

- Evaluate and aggregate contributor inputs to produce final decision
  - Majority vote
  - Discard inputs from low-trusted contributors (e.g. Hsueh et al. (2009))
  - MACE: a) identify which annotators are trustworthy and b) predict the correct underlying labels (Hovy 2013)
- Merge individual units from the microtasks (e.g. sentences) into complete documents, including all crowdsourced markup
- Tune the expert-created “gold” standard based on annotator feedback
  - Gold standard test questions often contain ambiguities and errors
  - Crowd has a broader knowledge-base than a few experts

Contributor 21271141: "GWTDT - Girl With The Dragon Tattoo is a film, therefore a product as it was made for sale." (0 Votes)

  - These are a great opportunity to train workers and amend expert data
  - Better gold data means better output quality, for the same cost
- To facilitate reuse, deliver the corpus in a widely used format, such as XCES, CONLL, GATE XML

Legal and Ethical Issues

1. Acknowledging the Crowd‘s contribution
   S. Cooper, [other authors], and **Foldit players**: Predicting protein structures with a multiplayer online game. Nature, 466(7307):756-760, 2010.

2. Ensuring privacy and wellbeing
   1. Mechnised labour criticised for low wages, lack of worker rights
   2. Majority of workers rely on microtasks as main income source
   3. Prevent prolonged use & user exploitation (e.g. daily caps)

3. Licensing and consent
   1. Some clearly state the use of Creative Common licenses
   2. General failure to provide informed consent information

Example: CF Instructions

Finding location names in text

Instructions ▲

In each sentence below, mark any names that are locations (e.g. **France**). Don't mark locations that don't have their own name.

There may be no locations in the sentence at all - that's OK.

Examples:

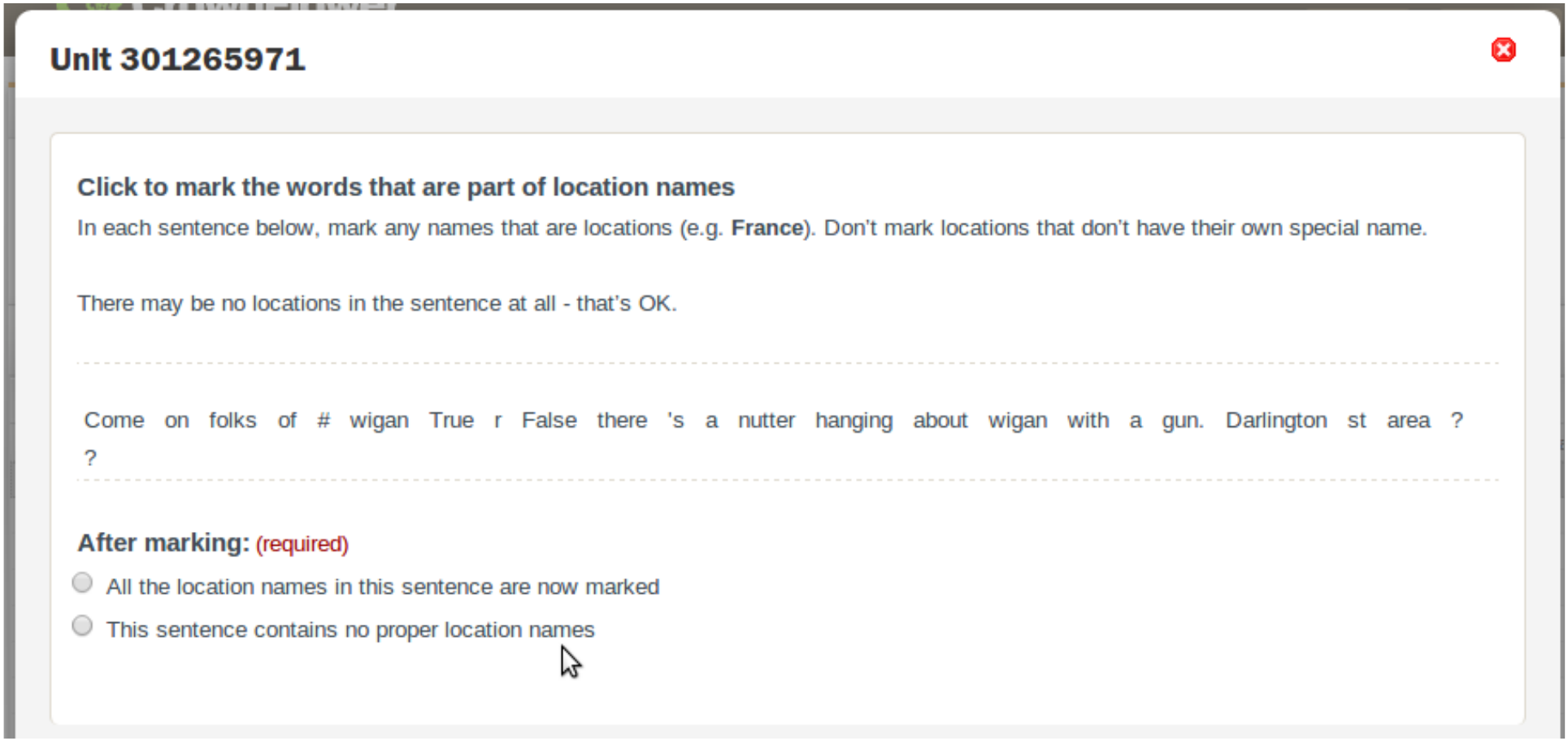
"There was a celebration in **London**"
correct - London is a location name

"The **room** is empty"
wrong, because room isn't the name of a particular location

"We traveled to **Spain** and had a great time **there**"
Only mark the location names, not words that just refer to it

"The award went to **Chelsea** Clinton"
wrong, because here Chelsea is a person

Example: CF Marking Locations in tweets



Example: CF Locations selected

Unit 301265971

Click to mark the words that are part of location names

In each sentence below, mark any names that are locations (e.g. **France**). Don't mark locations that don't have their own special name.

There may be no locations in the sentence at all - that's OK.

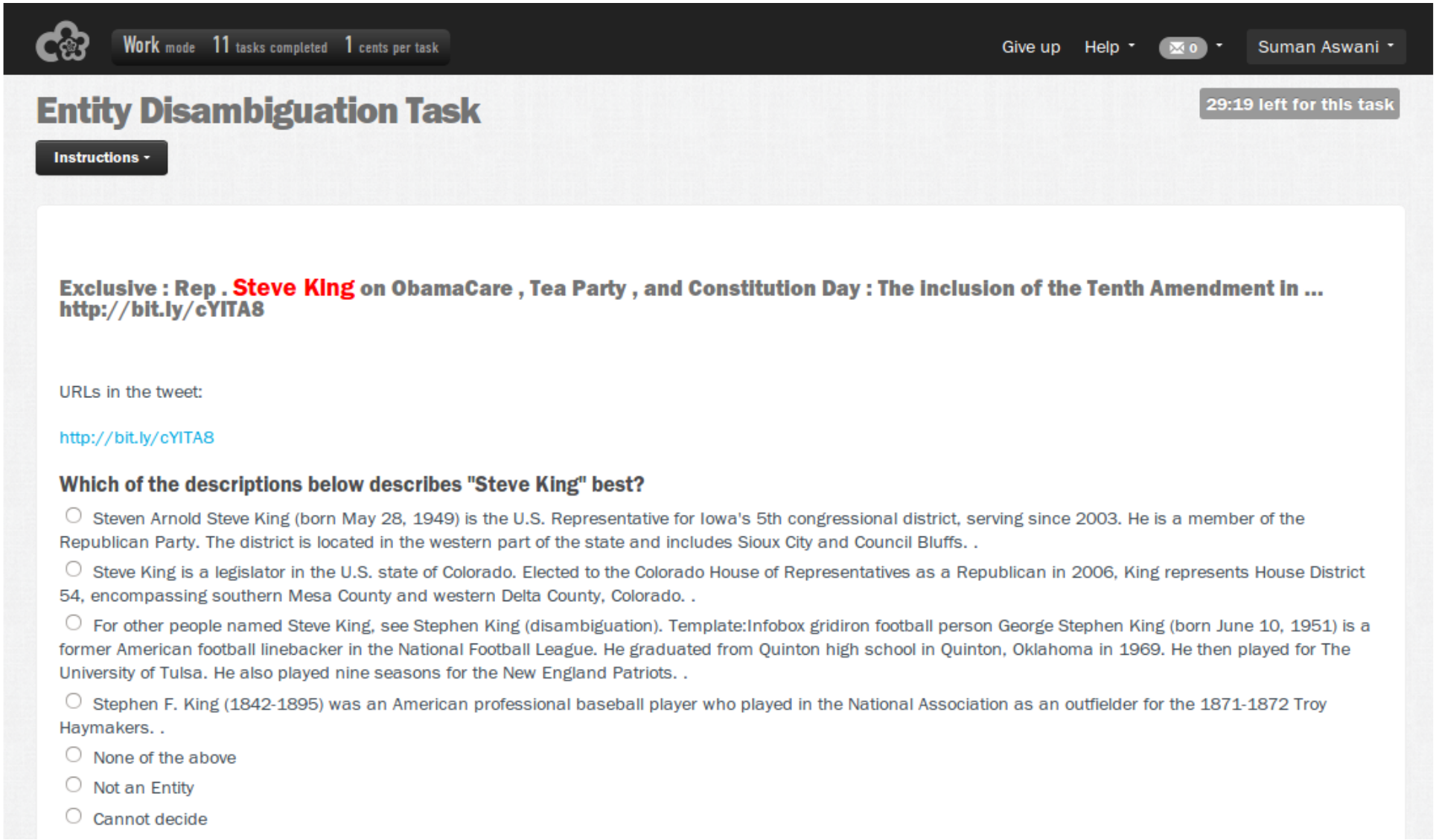
Come on folks of # wigan True r False there 's a nutter hanging about wigan with a gun. Darlington st area ? ?

After marking: (required)

- (•) All the location names in this sentence are now marked
- ( ) This sentence contains no proper location names

University of Sheffield,
NLP

Example 2: Entity Linking Annotation in CF



How to do it: The Laborious Way

- Export linguistic data as CSV file and load up into CrowdFlower
- Create instructions as HTML
- Customise the annotation UI (e.g. we had to use JavaScript for NE selection)
- Select how many judgments per micro-task and any restrictions on the annotators (e.g. country of origin)
- Test it and revisit any of the above, as needed
- Launch it and collect the data
- Download the results and put together the corpus
- Adjudicate

How to do it: The Easy Way

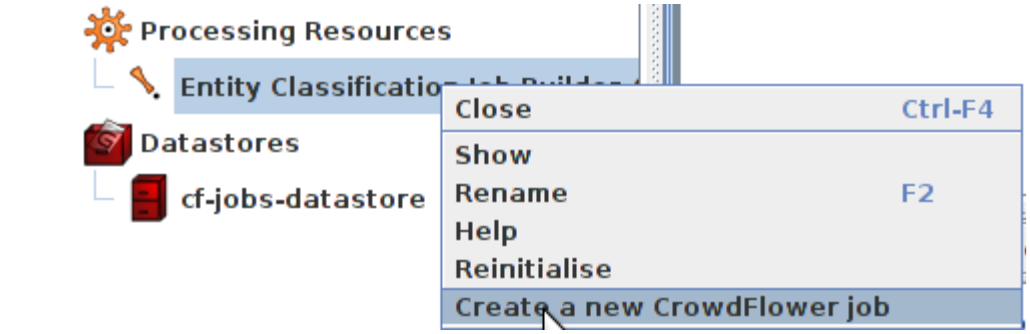
- Use the GATE Crowdsourcing plugin (release 8 onwards)
  - https://gate.ac.uk/wiki/crowdsourcing.html
- Transforms automatically texts with GATE annotations into CF jobs
- Generates the CF User Interface (based on templates)
- Researcher then checks and runs the project in CF
- On completion, the plugin imports automatically the results back into GATE, aligning to sentences and representing the multiple annotators
- To use, from the Plugin manager, load the Crowd_Sourcing plugin

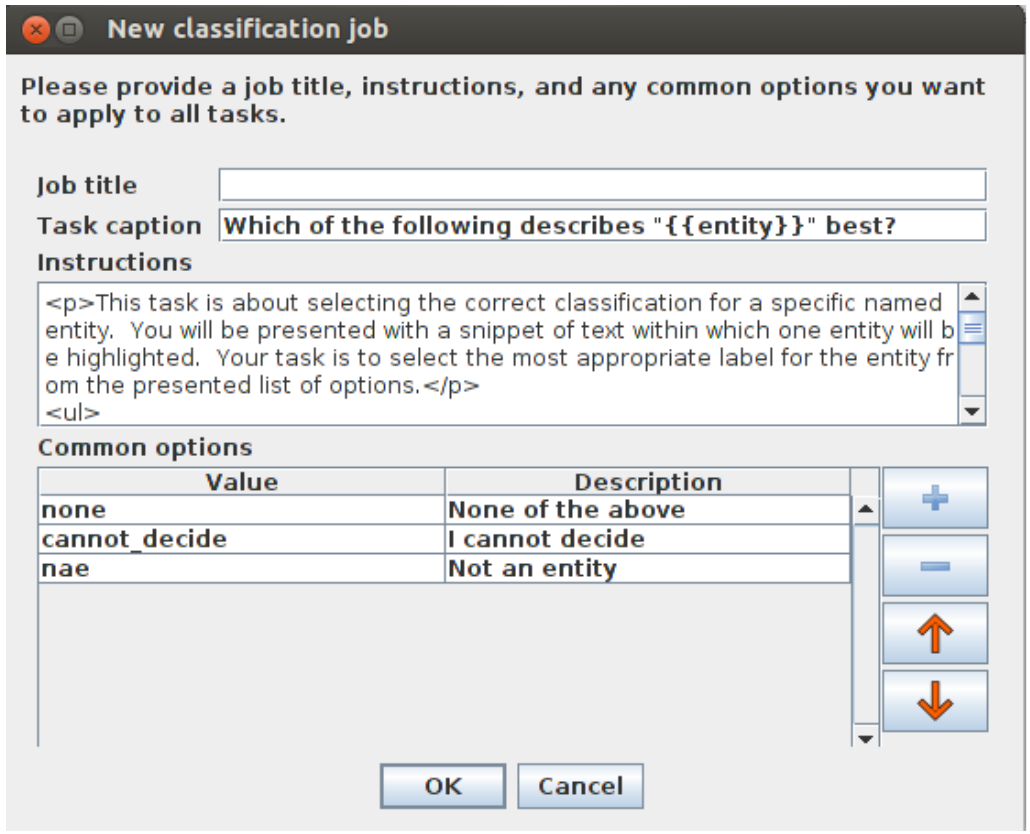
Crowd_Sourcing Plugin: Terminology

- *A job* - a single end-to-end crowdsourcing process. Holds a number of units of work
- *A unit* - single item of work. CrowdFlower presents several units at a time to the user as a single task, and users are paid for each task they successfully complete
- *A gold unit* - the correct answer is known in advance.
    - When a job includes gold units, CrowdFlower includes one gold unit in each task but does not tell the user which one it is, and if they get the gold unit wrong then the whole task is discarded.
    - You can track users' performance through the CrowdFlower platform and ignore results from users who get too many gold units wrong.

GATE Crowdsourcing Overview (1)

- Choose a job builder
  - Classification
  - Annotation
    (Sequence Selection)

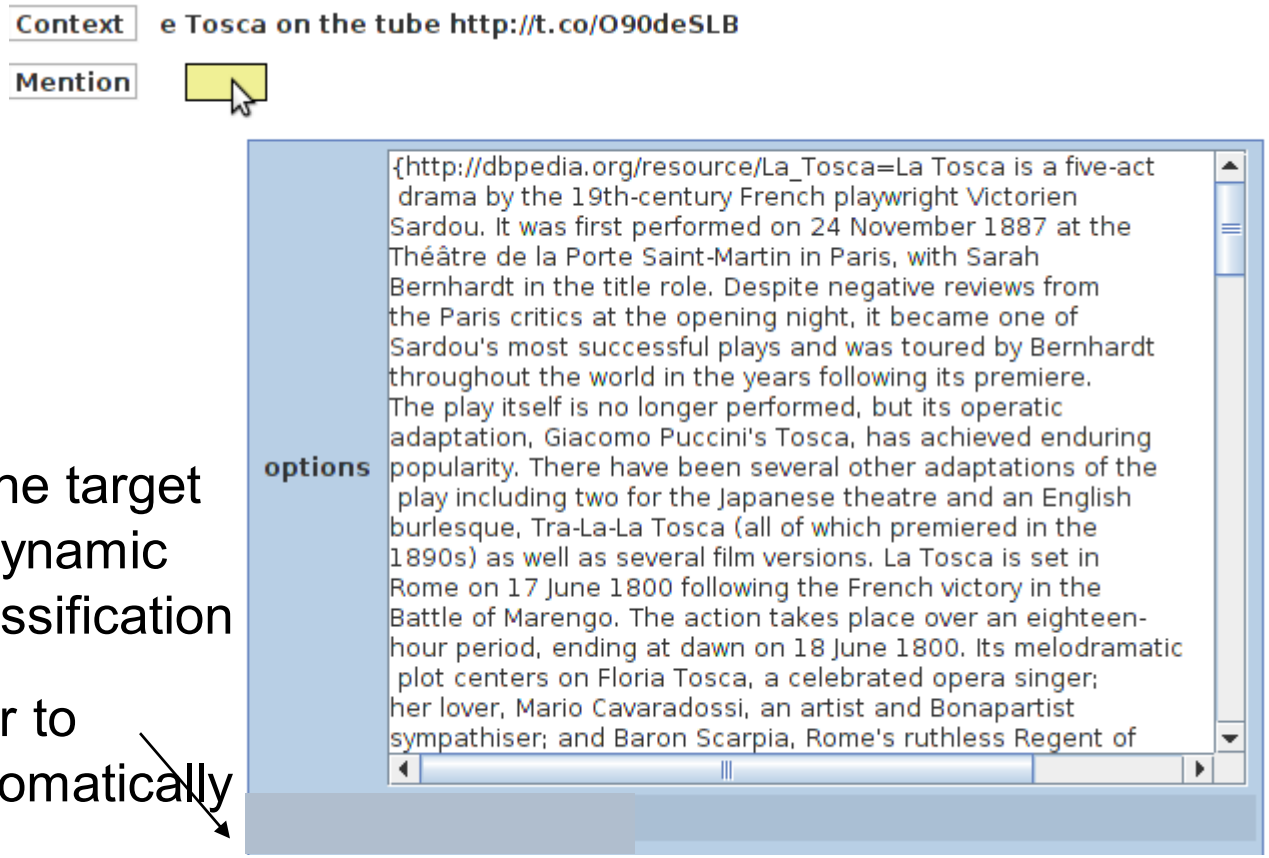




- Configure the corresponding user interface and provide the task instructions

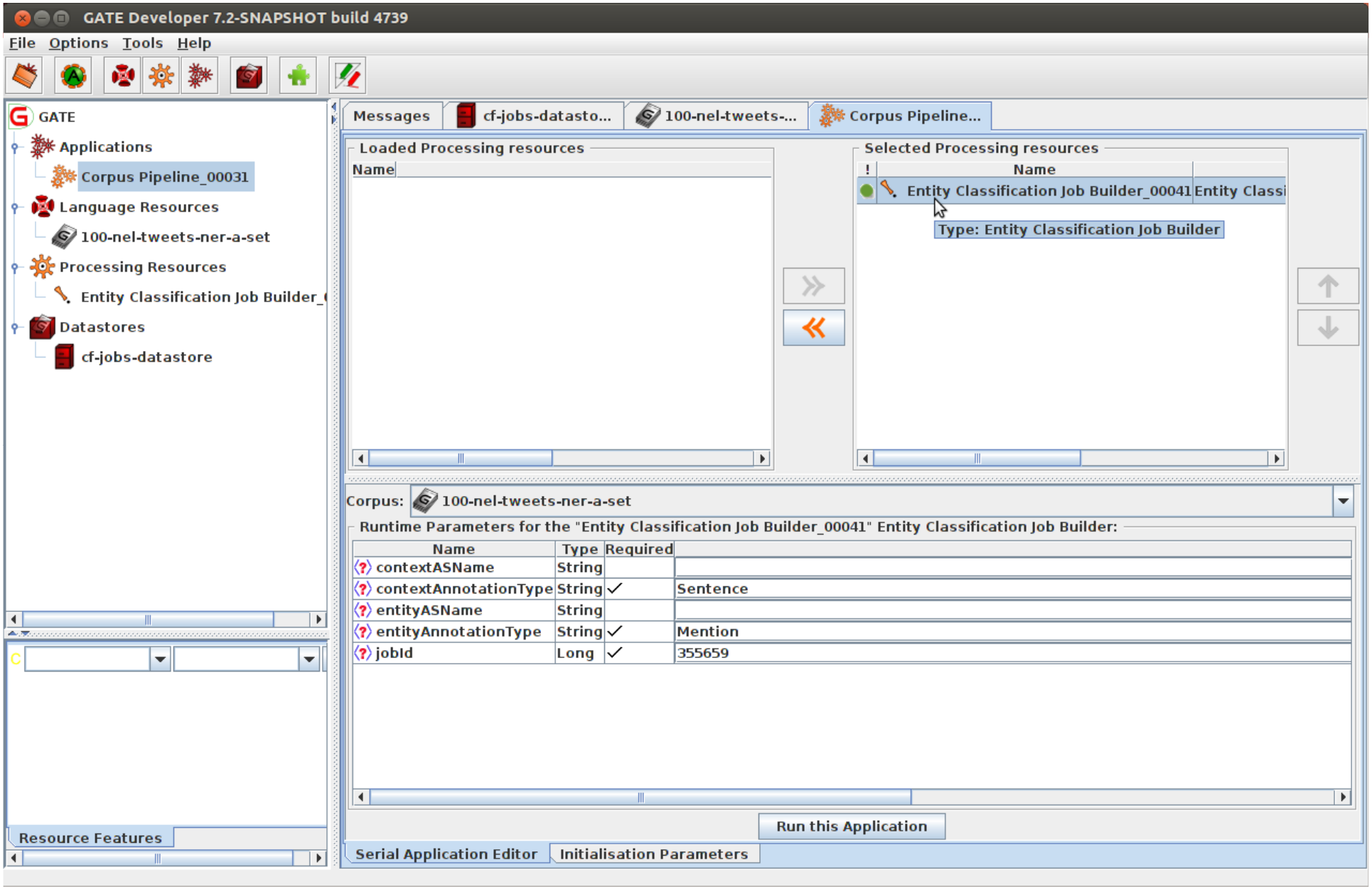
GATE Crowdsourcing Overview (2)

- Pre-process the corpus with TwitIE/ANNIE, e.g.
  - Tokenisation
  - POS tagging
  - Sentence splitting
  - NE recognition
- Save to a datastore
- Create automatically the target annotations and any dynamic values required for classification
- Execute the job builder to upload units to CF automatically

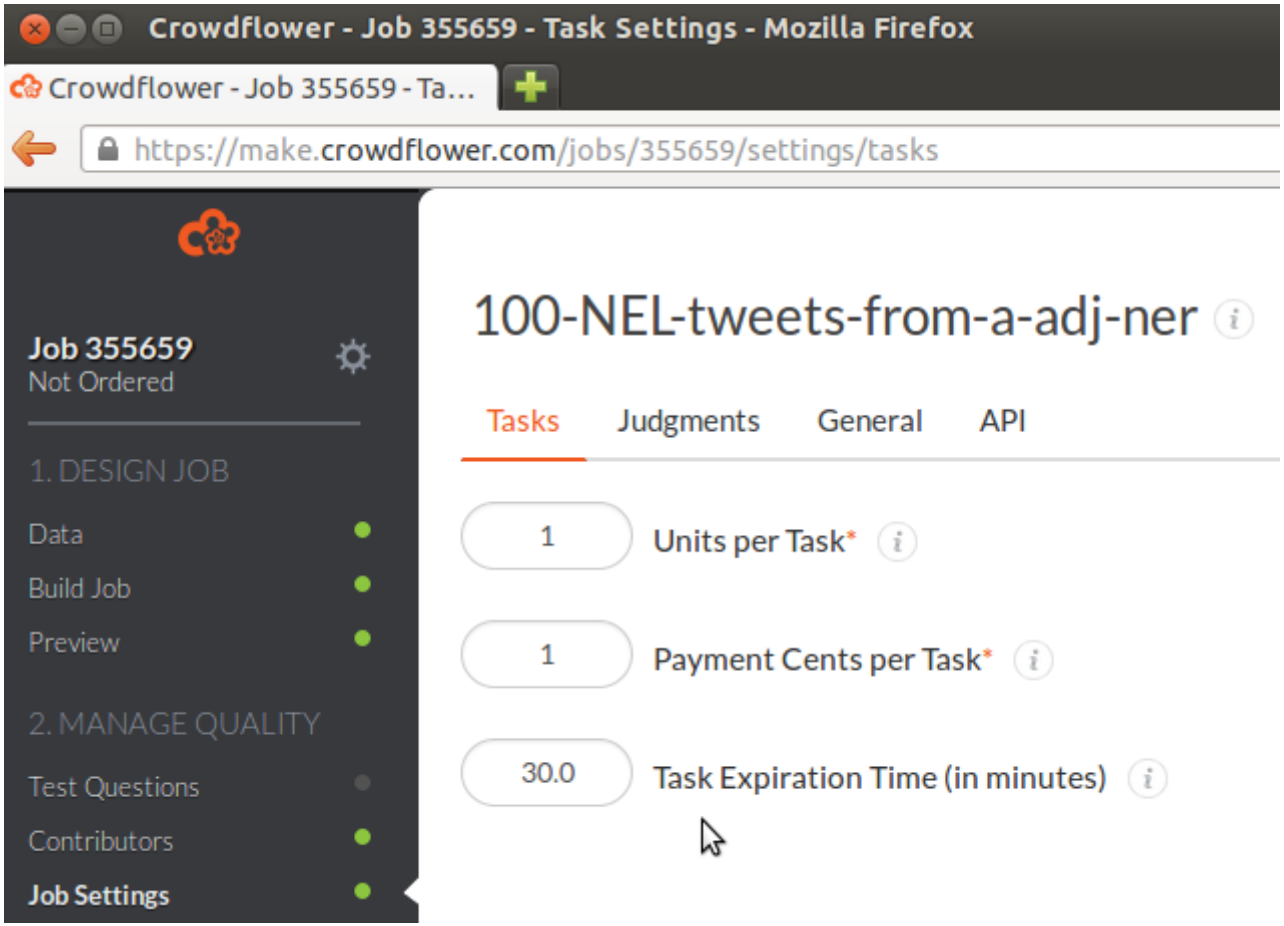


University of Sheffield,
NLP

GATE Crowdsourcing Overview (3)

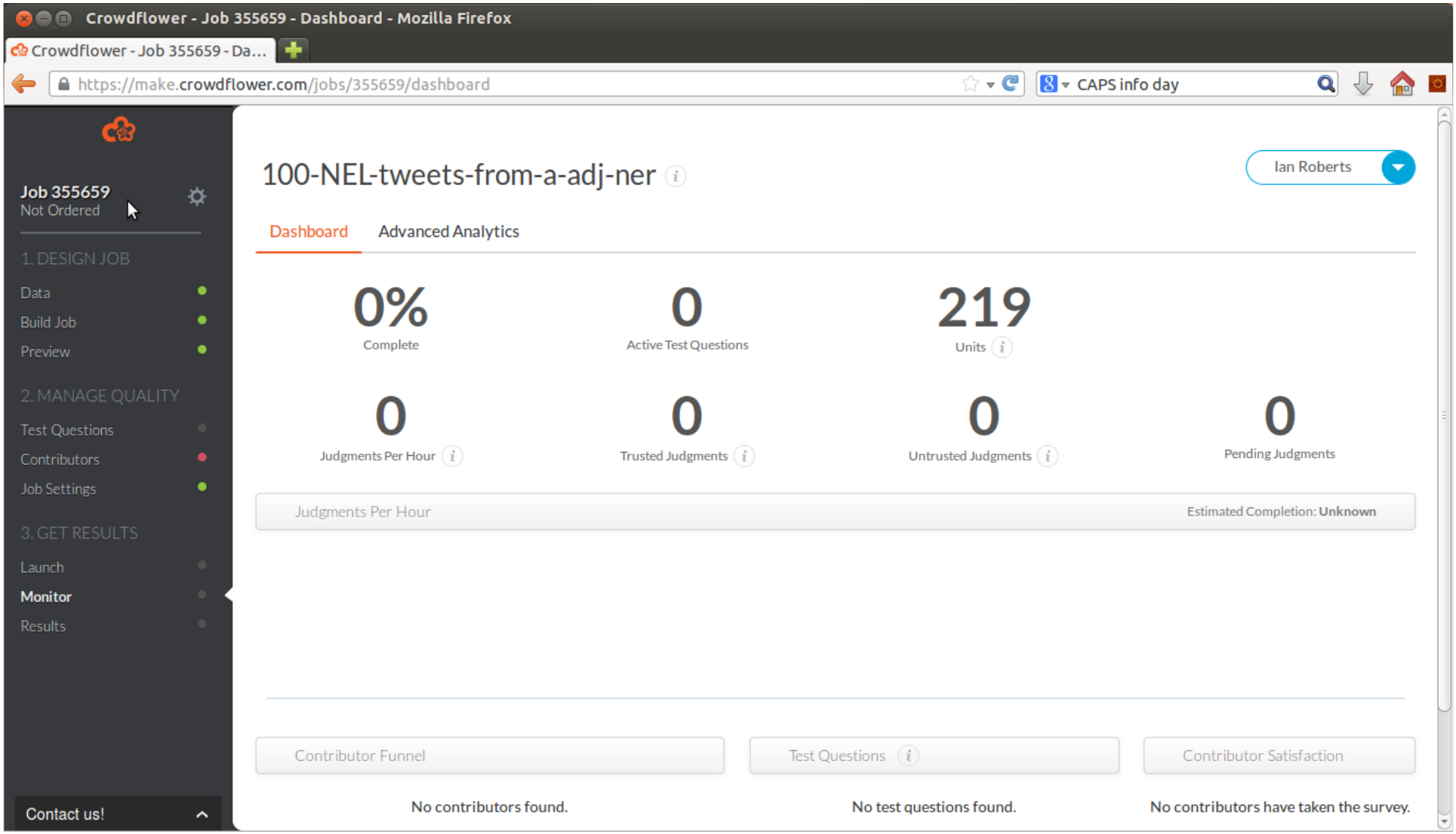


Configure and execute the job in CF



Gold data units can also be uploaded from GATE, so CF controls quality

CF Job Overview



Hands On: Classify named entities in CF

- Open http://bit.ly/1Td4lvf
- Login to CrowdFlower, as required
- Read the instructions and spend a few minutes annotating
- Make a note of any questions/issues you encounter
- Let's discuss them

Home work: Create a tweet classification CF job

- The aim is to crowdsource whether a set of tweets have positive/negative/neutral sentiment (i.e. classification job)
- Register with CrowdFlower for an API key
- Unpack hands-on-crowdsourcing.zip
- Load Datastore (sample-classification-ds) from within the hands-on
- Load the corpus from that datastore in GATE Developer
- Create an Entity Classification Job builder and give it your API key
- Right click on the Job builder/Create New CrowdFlower job
- Give it a job title, modify task captions and instructions to explain the sentiment classification task, and change the categories accordingly. You may keep none and cannot decide or remove them. Make sure the newly added classes are saved properly in the dialogue box

Home work (2)

- Add the Job builder PR to a new corpus pipeline
- Since we are classifying the entire tweets as pos/neg/neutral, specify text as the annotation type for both ContextAnnotationType and MentionAnnotationType (it is in the default set, so leave those blank)
- Set the skipExisting parameter to **false**
- Run the application
- Login to CrowdFlower, check and launch the job

Home work (3) – the job UI created

@Lobster JZombie I will survive global warming/climate change. #Evolution

Please indicate the sentiment expressed in this short text

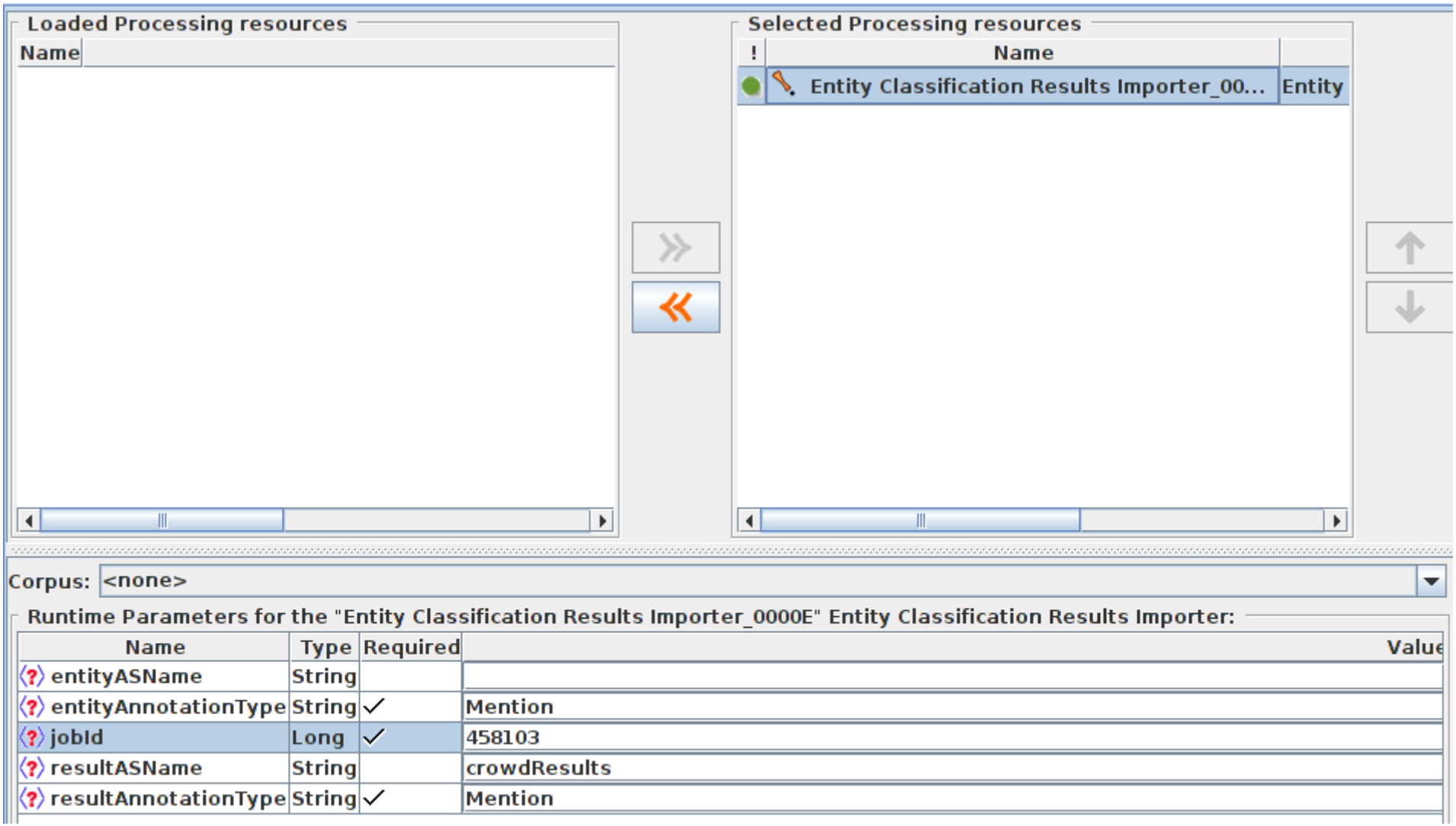
- ○ Neutral/No sentiment
- ○ Positive sentiment
- ○ Negative sentiment
- ○ None of the above
- ○ I cannot decide

Comment

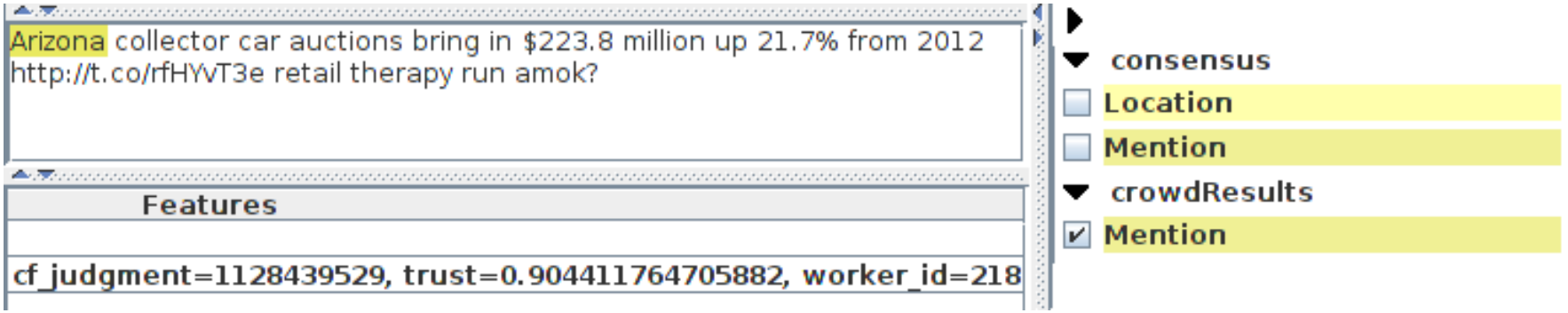
Importing CrowdFlower results into GATE

- Make sure the job is completed in CrowdFlower
- Load an Entity Classification or an Entity Annotation Results Importer, depending on what job you have created initially
- Add it to a corpus pipeline
- Provide the correct job ID by copying it from CrowdFlower
- Make sure the entityAnnotationType parameter has the correct value. For the tweet sentiment classification, for example, this would need to be changed to text
- Run the pipeline – it will iterate through the annotations and import the CF judgements automatically
- The results will be in the crowdResults set (unless you renamed it in the importer PR)

Importing CrowdFlower results (2)



Automatic CF Import into GATE

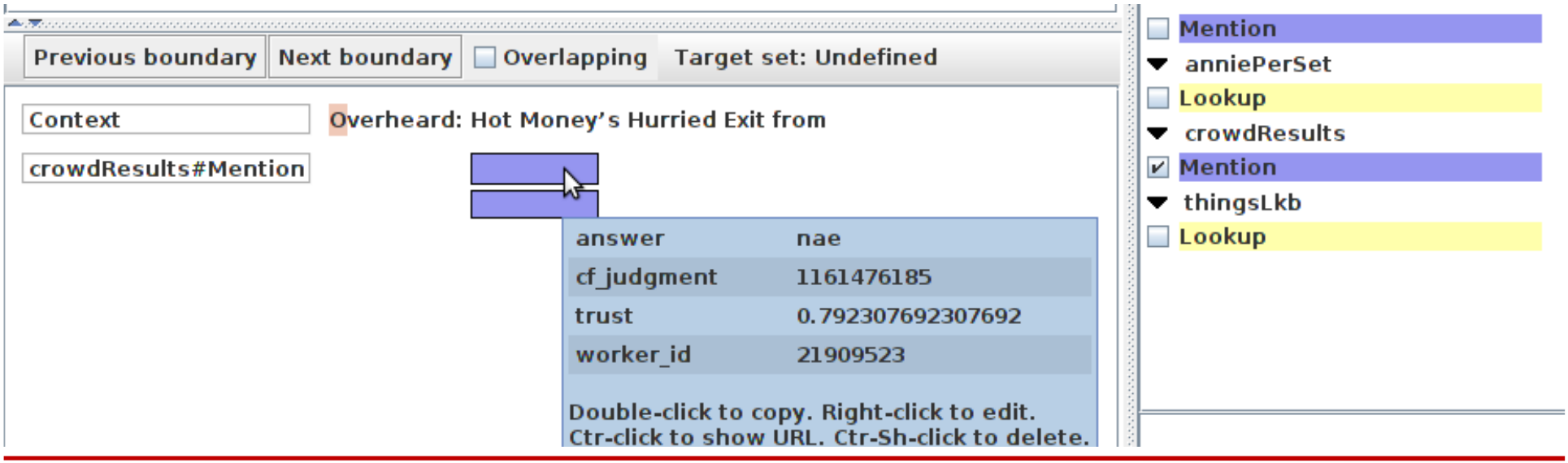


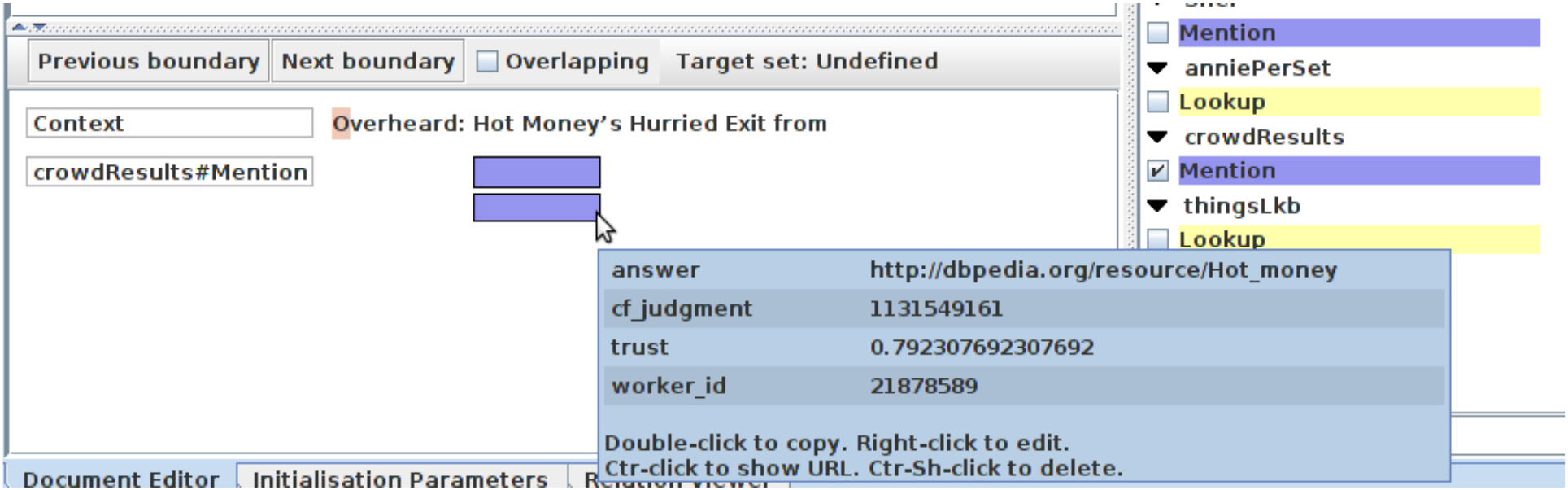
- Each CF judgement is imported back as a separate annotation with some metadata
- Adjudication can happen automatically (e.g. write a JAPE file to calculate majority vote) or manually (Annotation Stack editor)
- The resulting corpus is ready to use for experiments or can be exported out of GATE as XML/XCES

Manual adjudication: Annotation Stack

- Double click on each document, to view it
- Press the Annotations Stack button to show the editor
- Select Mention (or your target classification annotation type) in the crowdResults Annotation Set
- All judgements will be shown one underneath the other
- Press Previous/Next boundary buttons to navigate

Manual adjudication: Example





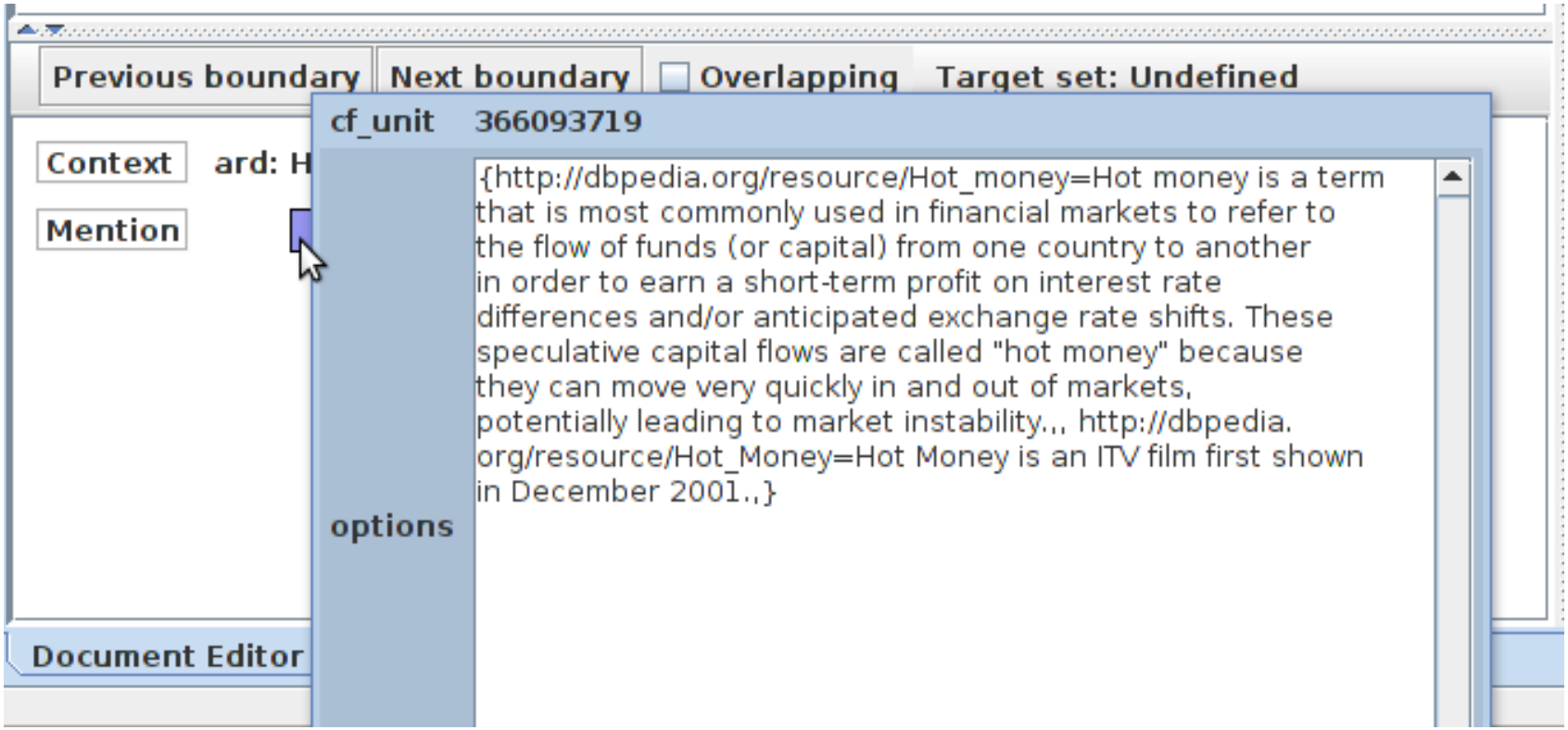
Manual adjudication: Annot. Stack (2)

- To adjudicate, double click on the annotation that you consider correct according to the annotation guidelines.
- This will copy the selected annotation into a new annotation set, together with all its features
- If more than 1 co-extensive annotation is correct, double click on just one of them (e.g. you don't want the gold standard to expect the system to annotate the same NE twice)
- Specify the target annotation set name, e.g. Key or consensus. You only need to do this once, then the same AS is used automatically
- Don't forget to save the document when you are finished

Manual adjudication: Hands on

- The task here is to disambiguate named entities by assigning them DBpedia URIs (values of the inst feature)
- From hands-on-crowdsourcing.zip, unpack the adjudication-exercise directory
- Create a corpus and populate it from that directory (11 docs)
- Double click several of them and try adjudicating the Mention annotations from the crowdResults annotation set
- Store the adjudicated annotations into the Key set
- For Mentions flagged as noe (not-an-entity), if you agree, then do not create a corresponding Mention in the Key set
- To see the choices shown in CF, enable the Annotations Stack to show also the Mention annotations from the default set

Hands on: Questions

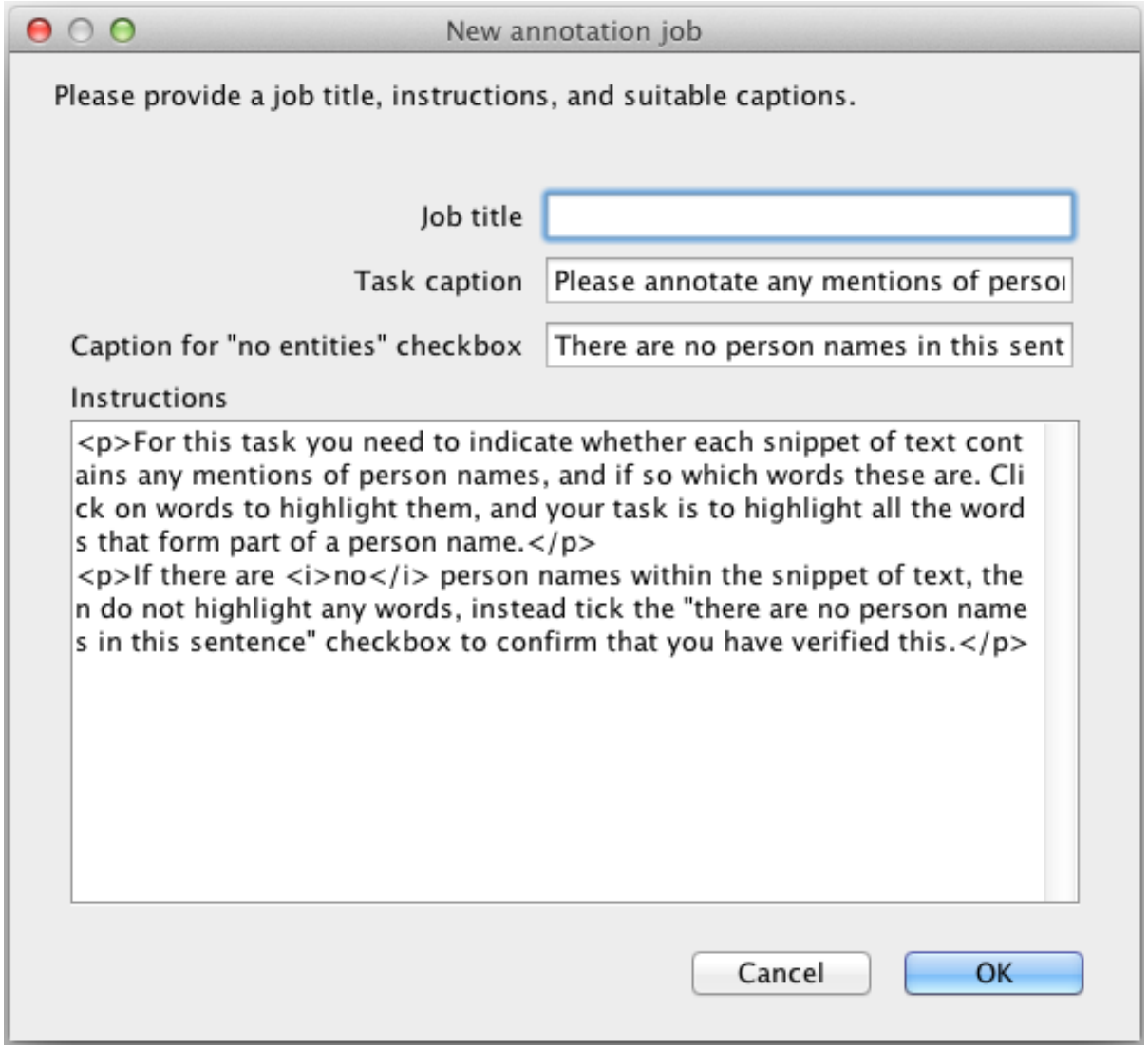


- In the last document, do you think Hot Money should be included as an entity with and URI or not?

Entity Annotation Jobs

- The “entity annotation” job builder and results importer PRs are for marking occurrences of named entities in plain text (or any sequence of tokens really)
- Assumptions:
  - Text is presented in short snippets (e.g. one sentence).
  - Each job focuses on one entity type. Annotating different entity types is done through running different jobs on the same corpus.
  - Entity annotations are whole tokens only, and there are no adjacent annotations (i.e. a contiguous sequence of marked tokens represents one target annotation)

Entity Annotation Jobs (2)



Entity Annotation Jobs (3)

Please annotate any mentions of person names in this sentence.

News of the approach follows last week 's £ 8 bn bid by a consortium of US finance groups for BT 's local telephone wires and could increase pressure on the group to consider a sell-off of infrastructure .

☑ There are no person names in this sentence

Please annotate any mentions of person names in this sentence.

The WestLB proposal is thought to have come in a meeting with Philip Hampton , BT 's finance director , several weeks ago .

Acknowledgements

Research partially supported by the uComp project (www.ucomp.eu). uComp receives the funding support of EPSRC EP/K017896/1, FWF 1097-N23, and ANR-12-CHRI-0003-03, in the framework of CHIST-ERA ERA-NET.

If using the GATE Crowdsourcing Plugin, please cite:

K. Bontcheva, I. Roberts, L. Derczynski, D. Rout. The GATE Crowdsourcing Plugin: Crowdsourcing Annotated Corpora Made Easy. Proceedings of the meeting of the European chapter of the Association for Computation Linguistics (EACL). 2014.

Bibliography

- G. Attardi. 2010. Phratris – A Phrase Annotation Game. In INSEMTIVES Game Idea Challenge
- C. Callison-Burch and M. Dredze. 2010a. Creating Speech and Language Data with Amazon's Mechanical Turk. In (Callison-Burch and Dredze, 2010b), pages 1–12.
- C. Callison-Burch and M. Dredze, editors. 2010b. Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk .
- T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. 2010. Annotating Named Entities in Twitter Data with Crowdsourcing. In Callison-Burch and Dredze (Callison-Burch and Dredze, 2010b), pages 80–88.
- K. Fort, G. Adda, and K.B. Cohen. 2011. Amazon Mechanical Turk: Gold Mine or Coal Mine? Computational Linguistics , 37(2):413 –420.
- Hoffmann, L. 2009. Crowd Control. Communications of the ACM , 52(3):16 –17.
- D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, E. Hovy. 2013. Learning Whom to Trust with MACE. Proc. NAACL
- P.Y. Hsueh, P. Melville, and V. Sindhwani. 2009. Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria. In Proc. of the Workshop on Active Learning for Natural Language Processing , pages 27–35.
- N. Lawson, K. Eustice, M. Perkowitz, and M. Yetisgen-Yildiz. 2010. Annotating Large Email Datasets for Named Entity Recognition with Mechanical Turk. In Callison-Burch and Dredze (Callison-Burch and Dredze, 2010b), pages 71–79.
- G. Parent and M. Eskenazi. 2011. Speaking to the Crowd: Looking at Past Achievements in Using Crowdsourcing for Speech and Predicting Future Challenges. In Proc. of INTERSPEECH , pages 3037– 3040.
- Poesio, M., U. Kruschwitz, J. Chamberlain, L. Robaldo, and L. Ducceschi. 2012. Phrase Detectives: Utilizing Collective Intelligence for Internet-Scale Language Resource Creation. Transactions on Interactive Intelligent Systems.
- A. Scharl, M. Sabou, S. Gindl, W. Rafelsberger, and A. Weichselbraun. 2012. Leveraging the wisdom of the crowds for the acquisition of multilingual language resources. Eight Int. Conf. on Language Resources and Evaluation Conference (LREC12) , pages 379–383.
- Snow, R. B. O'Connor, D. Jurafsky, and A. Y. Ng. 2008. Cheap and Fast—but is it Good?: Evaluating Non-Expert Annotations for Natural Language Tasks. In Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08) , pages 254– 263.
- Stede and C.R. Huang. 2012. Inter-operability and reusability: the science of annotation. Language Resources and Evaluation , 46:91–94. 10.1007/s10579-011-9164-x.
- L. von Ahn and L. Dabbish. 2008. Designing games with a purpose. Commun. ACM , 51(8):58–67
- A. Wang, C.D.V. Hoang, and M. Y. Kan. 2012. Perspectives on Crowdsourcing Annotations for Natural Language Processing. Language Resources and Evaluation
- O. F. Zaidan and C. Callison-Burch. 2011. Crowdsourcing Translation: Professional Quality from Non-Professionals. In Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL:HLT'11), pages 1220–1229.