



Crowdsourcing Social Media Corpora

Genevieve Gorrell





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
- New task types may require new annotations
- 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
- EACL 2014 paper: "[The GATE Crowdsourcing Plugin: Crowdsourcing Annotated Corpora Made Easy](#)" Bontcheva, Roberts, Derczynski, Rout
- Chapter in "Handbook of Linguistic Annotation", 2017: "[Crowdsourcing Named Entity Recognition and Entity Linking Corpora](#)" Bontcheva, Derczynski, Roberts



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), Figure Eight (FE, formerly CrowdFlower), Prolific
 - 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/FE 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)



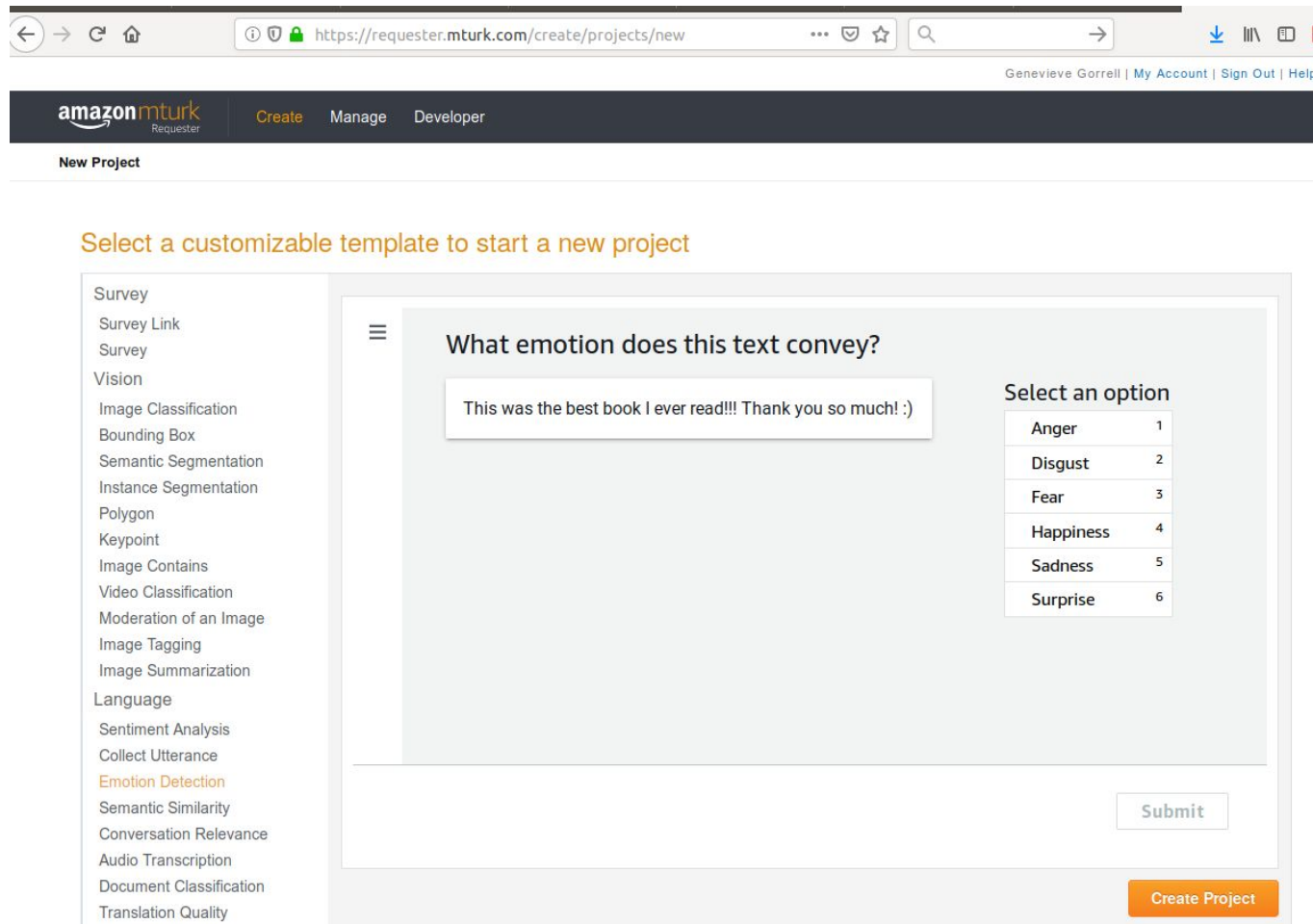


Paid for Crowdsourcing

- Contributors are extrinsically motivated through economic incentives
- Carry out micro-tasks in return for micro-payments
- Most NLP projects use crowdsourcing marketplaces: Amazon Mechanical Turk and Figure Eight
- Requesters post Human Intelligence Tasks (HITs) to a large population of micro-workers
- Challenges:
 - low quality output due to the workers' purely economic motivation
 - non-naive annotators and lack of annotator diversity
 - ethical issues (Fort et al., 2011)
 - Sequence effects (Mathur et al., 2017)

Amazon Mechanical Turk

- AMT is the biggest and best-known
- You can choose from various templates



https://requester.mturk.com/create/projects/new

Genevieve Gorrell | My Account | Sign Out | Help

amazonmturk Requester Create Manage Developer

New Project

Select a customizable template to start a new project

Survey
Survey Link
Survey
Vision
Image Classification
Bounding Box
Semantic Segmentation
Instance Segmentation
Polygon
Keypoint
Image Contains
Video Classification
Moderation of an Image
Image Tagging
Image Summarization
Language
Sentiment Analysis
Collect Utterance
Emotion Detection
Semantic Similarity
Conversation Relevance
Audio Transcription
Document Classification
Translation Quality

What emotion does this text convey?

This was the best book I ever read!!! Thank you so much! :)

Select an option

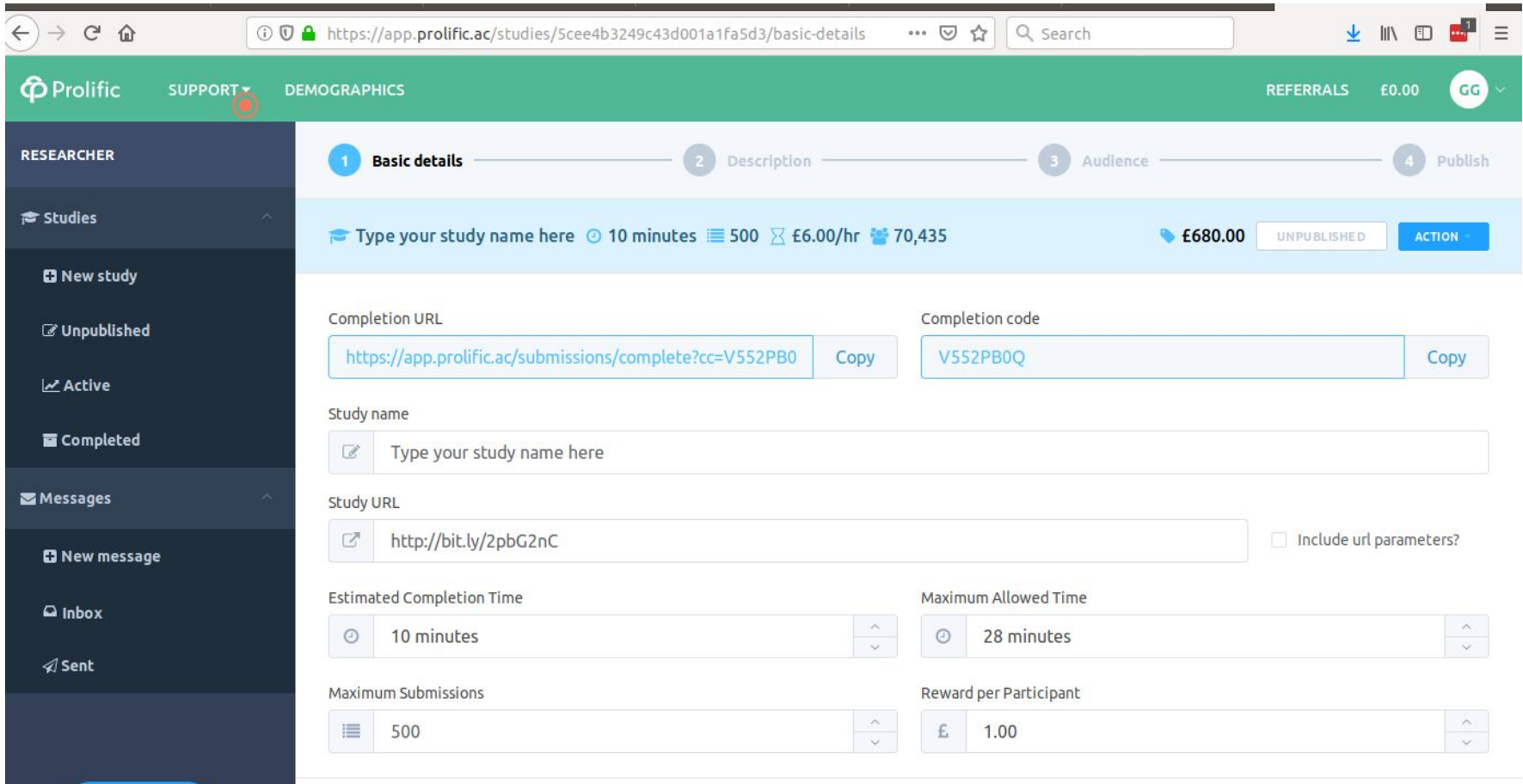
Anger	1
Disgust	2
Fear	3
Happiness	4
Sadness	5
Surprise	6

Submit

Create Project

Prolific Academic

- Prolific just manage the participants for you
- You have to construct your own study using other means



The screenshot displays the Prolific Academic web interface for configuring a study. The browser address bar shows the URL: <https://app.prolific.ac/studies/5cee4b3249c43d001a1fa5d3/basic-details>. The page features a green header with the Prolific logo, navigation links for SUPPORT and DEMOGRAPHICS, and user information including REFERRALS, £0.00, and a profile icon labeled GG.

The main content area is divided into a left sidebar and a main configuration panel. The sidebar, under the RESEARCHER heading, includes links for Studies, New study, Unpublished, Active, Completed, Messages, New message, Inbox, and Sent.

The main configuration panel shows a progress bar with four steps: 1 Basic details (active), 2 Description, 3 Audience, and 4 Publish. Below the progress bar, a summary bar displays: Type your study name here, 10 minutes, 500 submissions, £6.00/hr, 70,435 participants, £680.00 total, UNPUBLISHED status, and an ACTION button.

The configuration fields include:

- Completion URL: <https://app.prolific.ac/submissions/complete?cc=V552PB0> (with Copy button)
- Completion code: V552PB0Q (with Copy button)
- Study name: Type your study name here (with edit icon)
- Study URL: <http://bit.ly/2pbG2nC> (with edit icon and checkbox for "Include url parameters?")
- Estimated Completion Time: 10 minutes (with spinner)
- Maximum Allowed Time: 28 minutes (with spinner)
- Maximum Submissions: 500 (with spinner)
- Reward per Participant: £ 1.00 (with spinner)



Genre 2: Games with a purpose (GWAPs)

US08 Sentiment Quiz
Play Rankings Awards Feedback Help About

ECOresearch.net
Sentiment Quiz Award

Is the following a **negative**, **neutral** or **positive** statement about the candidate?

“ We are headed down a path that is certain to end in the destruction of our experiment in democracy. ”

[-] [-] [-] [+] [+] [!]

September 2008
843 players See All

1.	Fiorella	2458
2.	Michel	2241
3.	Birgit	2139
4.	Rose	1011
5.	Herti	930
...		
11.	Arno	101
12.	Guilherme	77
13.	You	65
14.	Lisa	61

Others currently playing

Election Monitor
Vote for your favorite candidate!

Barack Obama	John McCain	Cynthia McKinney

Status

4 Level

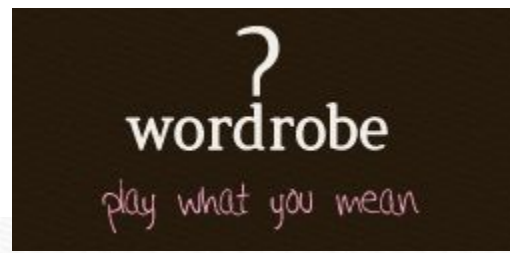
1514 13 12
65

Your current score is **65 points**. Invite your friends and earn 10% of the points they make!

Spread the Word
Tell your Friends!
You will earn **10% of your friends' points** after they accept your invitation! The calculation is recursive, so if they invite others you will even get more bonus points.

New Media MBA
www.modul.ac.at/nmt/mba

EDITED BOOK
The Geospatial Web
Geobrowsers, Social Software & the Web 2.0



Ranking (last 50 days)

1		Valerio		32150 points
2		wordrobe		5363 points
3		Aristotle		3998 points
4		sebb		3266 points
5		vincent		3028 points
6		arjanb		2495 points
7		EvaVanmassenhove		1308 points
8		furryfreak		1038 points



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)
- Challenges:
 - Designing appealing games and attracting a critical mass of players can be a challenge (Wang et al., 2012)

Genre 3: Altruistic Crowdsourcing



COMMUNITY

PROJECTS

ABOUT

SIGN IN

CREATE YOUR PROJECT

Become a *volunteer*.
Become a *researcher*.

We have hundreds of projects waiting for your help to achieve amazing goals.

SIGN UP AND BECOME A DIGITAL VOLUNTEER

It's free and 100% open sourced!

276 817

tasks done so far

5 755 251

pending tasks

209

projects

8 406

volunteers



Altruistic Crowdsourcing

- Intrinsic motivation can improve quality
- Project needs to be motivating/inspiring
- Good design is essential for increasing motivation

Workflow for Crowdsourcing Corpora

- 1a. Select NLP Problem and crowdsourcing genre
- 1b. Decompose NLP problem into tasks
- 1c. Design crowdsourcing task

- 2a. Collect and pre-process corpus
- 2b. Build or reuse annotator and management interfaces
- 2c. Run pilot studies

- 3a. Recruit and screen contributors
- 3b. Train, profile and retain contributors
- 3c. Manage and monitor crowdsourcing tasks

- 4a. Evaluate and aggregate annotations
- 4b. Evaluate overall corpus characteristics

1. Project Definition

2. Data and UI Preparation

3. Running the Project

4. Corpus Delivery





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 texts with nothing to annotate from texts that have not been annotated



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/FE and crowdcrafting
 - Significant investment for GWAPs or other custom interfaces
- Run bigger pilot studies with volunteers to test everything and collect gold units for quality control later



Step 3: Running the Crowdsourcing Project

Monitor

STATUS **Finished** [View Results](#)

You have finalized rows that can be converted into test questions [Start](#)

[Dashboard](#) [Alerts*](#) [Advanced Analytics](#) [Contributors](#)

100%

Complete ⓘ

\$50

Cost ⓘ

72

Active Test Questions ⓘ

300

Rows ⓘ

0

Judgments By Hour ⓘ

1,500

Trusted Judgments ⓘ

586

Untrusted Judgments ⓘ

0

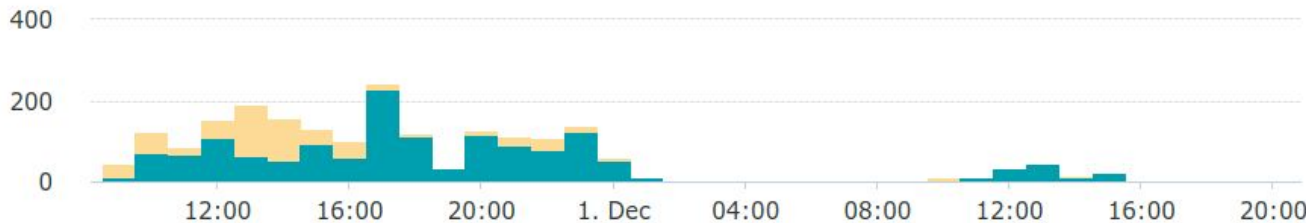
Pending Judgments ⓘ

- [Settings](#)
- [Preview](#)
- [Copy ▾](#)

- READ THE DOCS
- [Dashboard Guide](#)
 - [Troubleshooting Guide](#)
 - [Auto-paused Jobs](#)
 - [Converting Finalized Rows](#)

Judgments By Hour

Estimated Completion: Dec 1, 8:15 pm





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
 - Avoids all data coming from small number of contributors
 - Needs less gold data as acquired data can be leveraged



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 et al 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. Mechanised 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: FE 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: FE Marking Locations in tweets

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





Example: FE 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

Example 2: Entity Linking Annotation in FE



Work mode 11 tasks completed 1 cents per task

Give up

Help ▾

✉ 0 ▾

Suman Aswani ▾

Entity Disambiguation Task

29:19 left for this task

Instructions ▾

Exclusive : Rep . **Steve King on ObamaCare , Tea Party , and Constitution Day : The Inclusion of the Tenth Amendment In ...**
<http://bit.ly/cYITAB>

URLs in the tweet:

<http://bit.ly/cYITAB>

Which of the descriptions below describes "Steve King" best?

- Steven Arnold Steve King (born May 28, 1949) is the U.S. Representative for Iowa's 5th congressional district, serving since 2003. He is a member of the Republican Party. The district is located in the western part of the state and includes Sioux City and Council Bluffs. .
- Steve King is a legislator in the U.S. state of Colorado. Elected to the Colorado House of Representatives as a Republican in 2006, King represents House District 54, encompassing southern Mesa County and western Delta County, Colorado. .
- For other people named Steve King, see Stephen King (disambiguation). Template:Infobox gridiron football person George Stephen King (born June 10, 1951) is a former American football linebacker in the National Football League. He graduated from Quinton high school in Quinton, Oklahoma in 1969. He then played for The University of Tulsa. He also played nine seasons for the New England Patriots. .
- Stephen F. King (1842-1895) was an American professional baseball player who played in the National Association as an outfielder for the 1871-1872 Troy Haymakers. .
- None of the above
- Not an Entity
- Cannot decide



How to do it: The Laborious Way

- Export linguistic data as CSV file and load up into Figure Eight
- Create instructions as HTML
- Customise the annotation UI (e.g. we had to use JavaScript for NE selection)
- After completion, download the results and put together the corpus
- Adjudicate



How to do it: The Easy Way

- Use the GATE Crowdsourcing plugin
 - <https://gate.ac.uk/wiki/crowdsourcing.html>
 - https://github.com/GateNLP/gateplugin-Crowd_Sourcing
- Automatically transforms texts with GATE annotations into FE jobs
- Generates the FE User Interface (based on 2 templates; disambiguation or entity annotation)
- Researcher then sets up annotator restrictions, checks and runs the project in FE
- On completion, the plugin automatically imports the results back into GATE, aligning to sentences and representing the multiple annotators
- To use, from the Plugin manager, load the “Crowd Sourcing” plugin

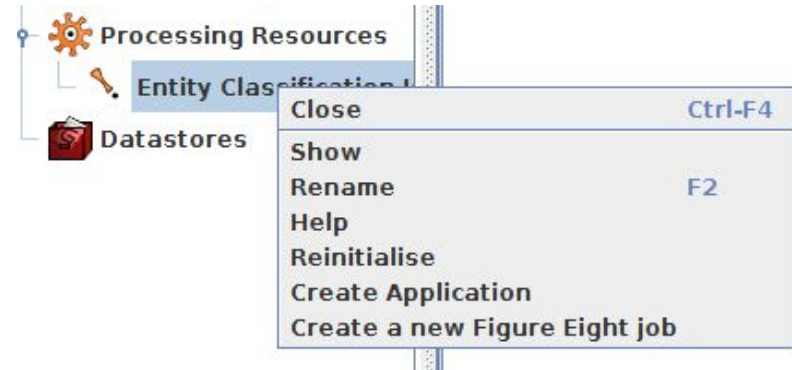


Crowd_Sourcing Plugin: Terminology

- *A job* - a single end-to-end crowdsourcing process. Holds a number of units of work
- *A unit* - single item of work. Figure Eight 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, Figure Eight 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 Figure Eight 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)



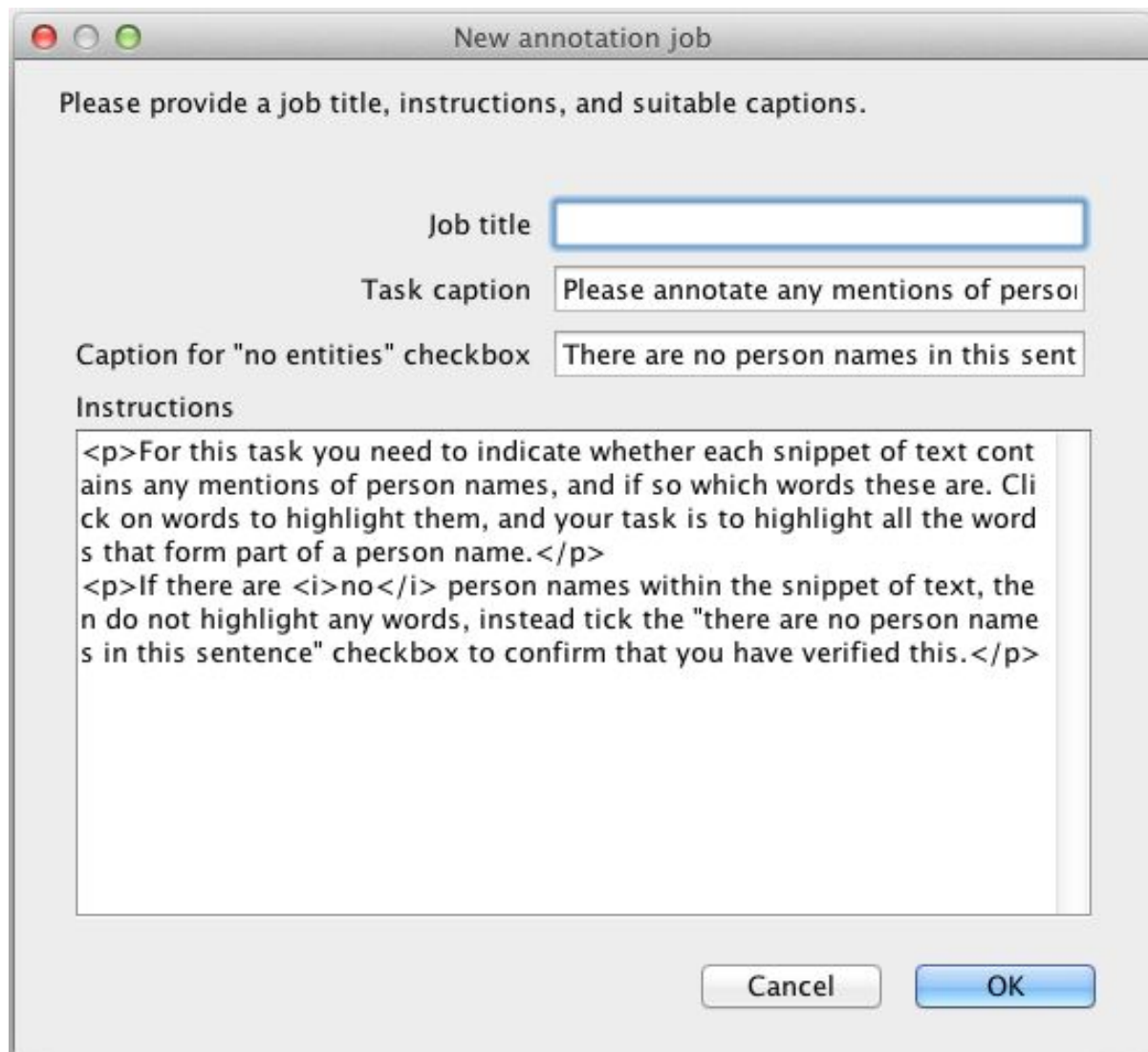
- Configure the corresponding user interface and provide the task instructions



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)

A screenshot of a dialog box titled "New annotation job" from the GATE software. The dialog box has a standard Mac OS window title bar with red, yellow, and green buttons. The main content area contains the following elements:

- A prompt: "Please provide a job title, instructions, and suitable captions."
- A "Job title" label followed by an empty text input field.
- A "Task caption" label followed by a text input field containing the text "Please annotate any mentions of person".
- A "Caption for 'no entities' checkbox" label followed by a text input field containing the text "There are no person names in this sent".
- An "Instructions" label followed by a large text area containing two paragraphs of text:

<p>For this task you need to indicate whether each snippet of text contains any mentions of person names, and if so which words these are. Click on words to highlight them, and your task is to highlight all the words that form part of a person name.</p>
<p>If there are <i>no</i> person names within the snippet of text, then do not highlight any words, instead tick the "there are no person names in this sentence" checkbox to confirm that you have verified this.</p>
- At the bottom, there are two buttons: "Cancel" and "OK".

Entity Annotation Jobs (3)

Please annotate any mentions of person names in this sentence.

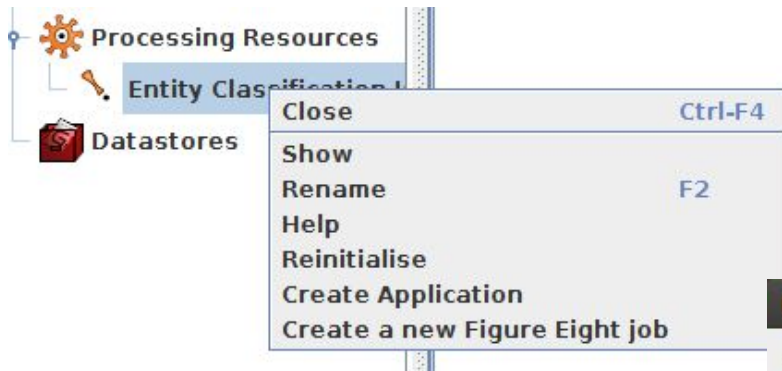
News of the approach follows last week 's £ 8 bn bid by a consortium of US finance groups for BT 's local telephone wires and could increase pressure on the group to consider a sell-off of infrastructure .

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 .

Classification Jobs



New classification job

Please provide a job title, instructions, and any common options you want to apply to all tasks.

Job title

Task caption

Instructions

Common options

Value	Description
none	None of the above
cannot_decide	I cannot decide
nae	Not an entity

OK Cancel

Classification Jobs

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

Context e Tosca on the tube <http://t.co/O90deSLB>

Mention 


options

```
{http://dbpedia.org/resource/La_Tosca=La Tosca is a five-act
drama by the 19th-century French playwright Victorien
Sardou. It was first performed on 24 November 1887 at the
Théâtre de la Porte Saint-Martin in Paris, with Sarah
Bernhardt in the title role. Despite negative reviews from
the Paris critics at the opening night, it became one of
Sardou's most successful plays and was toured by Bernhardt
throughout the world in the years following its premiere.
The play itself is no longer performed, but its operatic
adaptation, Giacomo Puccini's Tosca, has achieved enduring
popularity. There have been several other adaptations of the
play including two for the Japanese theatre and an English
burlesque, Tra-La-La Tosca (all of which premiered in the
1890s) as well as several film versions. La Tosca is set in
Rome on 17 June 1800 following the French victory in the
Battle of Marengo. The action takes place over an eighteen-
hour period, ending at dawn on 18 June 1800. Its melodramatic
plot centers on Floria Tosca, a celebrated opera singer;
her lover, Mario Cavaradossi, an artist and Bonapartist
sympathiser; and Baron Scarpia, Rome's ruthless Regent of
```

GATE Crowdsourcing Overview (3)

GATE Developer 7.2-SNAPSHOT build 4739

File Options Tools Help



GATE

- Applications
 - Corpus Pipeline_00031
- Language Resources
 - 100-nel-tweets-ner-a-set
- Processing Resources
 - Entity Classification Job Builder_00041
- Datastores
 - cf-jobs-datastore

Messages cf-jobs-datasto... 100-nel-tweets-... Corpus Pipeline...

Loaded Processing resources

Name

Selected Processing resources

!	Name	
	Entity Classification Job Builder_00041	Entity Classi
	Type: Entity Classification Job Builder	

Corpus: 100-nel-tweets-ner-a-set

Runtime Parameters for the "Entity Classification Job Builder_00041" Entity Classification Job Builder:

Name	Type	Required	
contextASName	String		
contextAnnotationType	String	✓	Sentence
entityASName	String		
entityAnnotationType	String	✓	Mention
jobId	Long	✓	355659

Run this Application

Serial Application Editor Initialisation Parameters



Configure and execute the job in FE

A screenshot of the Crowdfunder web interface showing the task settings for a job. The browser title is "Crowdfunder - Job 355659 - Task Settings - Mozilla Firefox". The address bar shows the URL "https://make.crowdfunder.com/jobs/355659/settings/tasks". The main heading is "100-NEL-tweets-from-a-adj-ner". There are four tabs: "Tasks" (selected), "Judgments", "General", and "API". The "Tasks" tab contains three settings: "Units per Task*" with a value of 1, "Payment Cents per Task*" with a value of 1, and "Task Expiration Time (in minutes)" with a value of 30.0. A left sidebar shows the job details "Job 355659 Not Ordered" and a navigation menu with sections "1. DESIGN JOB" and "2. MANAGE QUALITY".

Crowdfunder - Job 355659 - Task Settings - Mozilla Firefox

Crowdfunder - Job 355659 - Ta... +

https://make.crowdfunder.com/jobs/355659/settings/tasks

Job 355659
Not Ordered

1. DESIGN JOB

Data ●

Build Job ●

Preview ●

2. MANAGE QUALITY

Test Questions ●

Contributors ●

Job Settings ●

100-NEL-tweets-from-a-adj-ner ⓘ

Tasks Judgments General API

1 Units per Task* ⓘ

1 Payment Cents per Task* ⓘ

30.0 Task Expiration Time (in minutes) ⓘ

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



FE Job Overview

Crowdflower - Job 355659 - Dashboard - Mozilla Firefox

Crowdflower - Job 355659 - Da...

https://make.crowdflower.com/jobs/355659/dashboard

CAPS info day

100-NEL-tweets-from-a-adj-ner Info Ian Roberts

Dashboard Advanced Analytics

0% Complete	0 Active Test Questions	219 Units Info	
0 Judgments Per Hour Info	0 Trusted Judgments Info	0 Untrusted Judgments Info	0 Pending Judgments

Judgments Per Hour Estimated Completion: Unknown

Contributor Funnel No contributors found.

Test Questions Info No test questions found.

Contributor Satisfaction No contributors have taken the survey.

Contact us!



Homework: Create a tweet classification FE job

- The aim is to crowdsource whether a set of tweets have positive/negative/neutral sentiment (i.e. classification job)
- Register with Figure Eight for an API key (new “customer” account—you need a second email address)
- 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 Figure Eight job
- Give it a job title, modify task captions and instructions to explain the sentiment classification task, and change the categories accordingly (pos/neg). You may keep “none” and “cannot decide” or remove them. Make sure the newly added classes are saved properly in the dialogue box



Homework (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 entityAnnotationType (it is in the default set, so leave those blank)
- Set the skipExisting parameter to **false**
- Run the application
- Login to Figure Eight
- Set the channels to internal only
- Check and launch the job
- See the bottom of the “Monitor” page for a sharable link.



Home work (3) – the job UI created

@LobsterJZombie 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



Hands On: Classify named entities in FE

- Open <https://tinyurl.com/tryannotation>
- Login to Figure Eight, as required
- Read the instructions and spend a few minutes annotating
- Make a note of any questions/issues you encounter
- Let's discuss them



Hands On: Games with a Purpose

- Play some of these:
 - <http://apetopia.visual-computing.com/>
 - <https://eternagame.org/> Bit more complicated
 - <http://guessthecorrelation.com/>
 - <http://anawiki.essex.ac.uk/phrasedetectives/> (trivial login)
 - <https://phylo.cs.mcgill.ca/play.html> Bit more complicated
 - <http://smorballgame.org/>



Importing Figure Eight results into GATE

- Make sure the job is completed in Figure Eight
- 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 Figure Eight
- 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 FE judgements automatically
- The results will be in the `crowdResults` set (unless you renamed it in the importer PR)

Importing Figure Eight results (2)

Loaded Processing resources

Name

Selected Processing resources

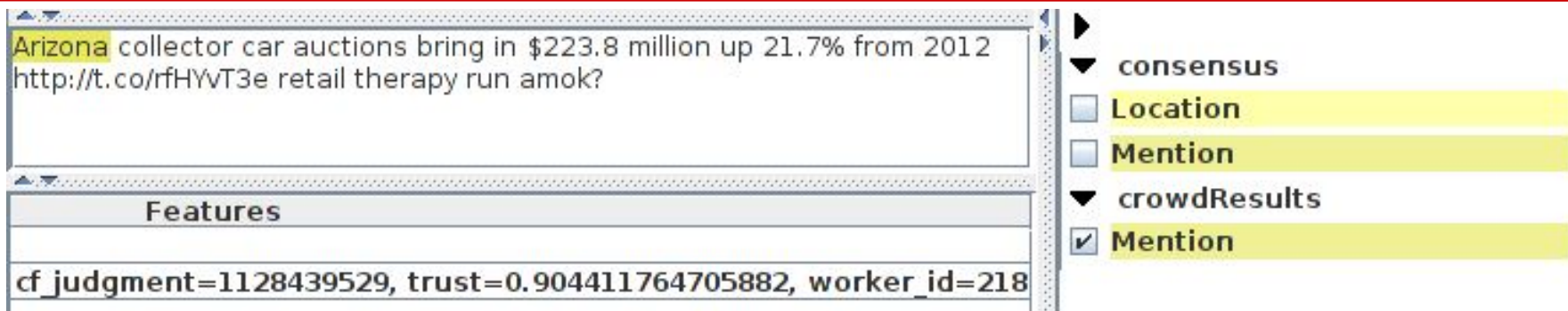
!	Name	
●	Entity Classification Results Importer_00...	Entity

Corpus: <none>

Runtime Parameters for the "Entity Classification Results Importer_0000E" Entity Classification Results Importer:

Name	Type	Required	Value
? entityASName	String		
? entityAnnotationType	String	✓	Mention
? jobId	Long	✓	458103
? resultASName	String		crowdResults
? resultAnnotationType	String	✓	Mention

Automatic FE Import into GATE

A screenshot of the GATE software interface. The top window displays a text snippet: "Arizona collector car auctions bring in \$223.8 million up 21.7% from 2012 http://t.co/rfHYVT3e retail therapy run amok?". Below this, a "Features" table is visible, with the first row containing the text "cf_judgment=1128439529, trust=0.904411764705882, worker_id=218". To the right, a sidebar shows a list of feature categories: "consensus" (expanded), "Location" (unchecked), "Mention" (unchecked), "crowdResults" (expanded), and "Mention" (checked). The "Location" and "Mention" items under "consensus" are highlighted in yellow.

- Each FE judgement is imported back as a separate annotation with some metadata
- Adjudication can happen automatically using PR provided or a custom script, or manually via the 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

Previous boundary Next boundary Overlapping Target set: Undefined

Context Overheard: Hot Money's Hurried Exit from

crowdResults#Mention

answer	nae
cf_judgment	1161476185
trust	0.792307692307692
worker_id	21909523

Double-click to copy. Right-click to edit.
Ctr-click to show URL. Ctr-Sh-click to delete.

- Mention
- ▼ anniePerSet
- Lookup
- ▼ crowdResults
- Mention
- ▼ thingsLkb
- Lookup

Previous boundary Next boundary Overlapping Target set: Undefined

Context Overheard: Hot Money's Hurried Exit from

crowdResults#Mention

answer	http://dbpedia.org/resource/Hot_money
cf_judgment	1131549161
trust	0.792307692307692
worker_id	21878589

Double-click to copy. Right-click to edit.
Ctr-click to show URL. Ctr-Sh-click to delete.

- Mention
- ▼ anniePerSet
- Lookup
- ▼ crowdResults
- Mention
- ▼ thingsLkb
- Lookup

Document Editor Initialisation Parameters Reasoner



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



Automatic Adjudication

- Annotations can also be adjudicated automatically, by “voting” between annotators.
- Use the two Majority-vote consensus builder PRs for this.
- We can set a minimum threshold for agreement
 - For example, refusing to accept an answer on which fewer than two out of three annotators agreed.
- Disputed judgments can then either be classified by hand, or fed back to Figure Eight as a new job.



Manual adjudication: Hands on

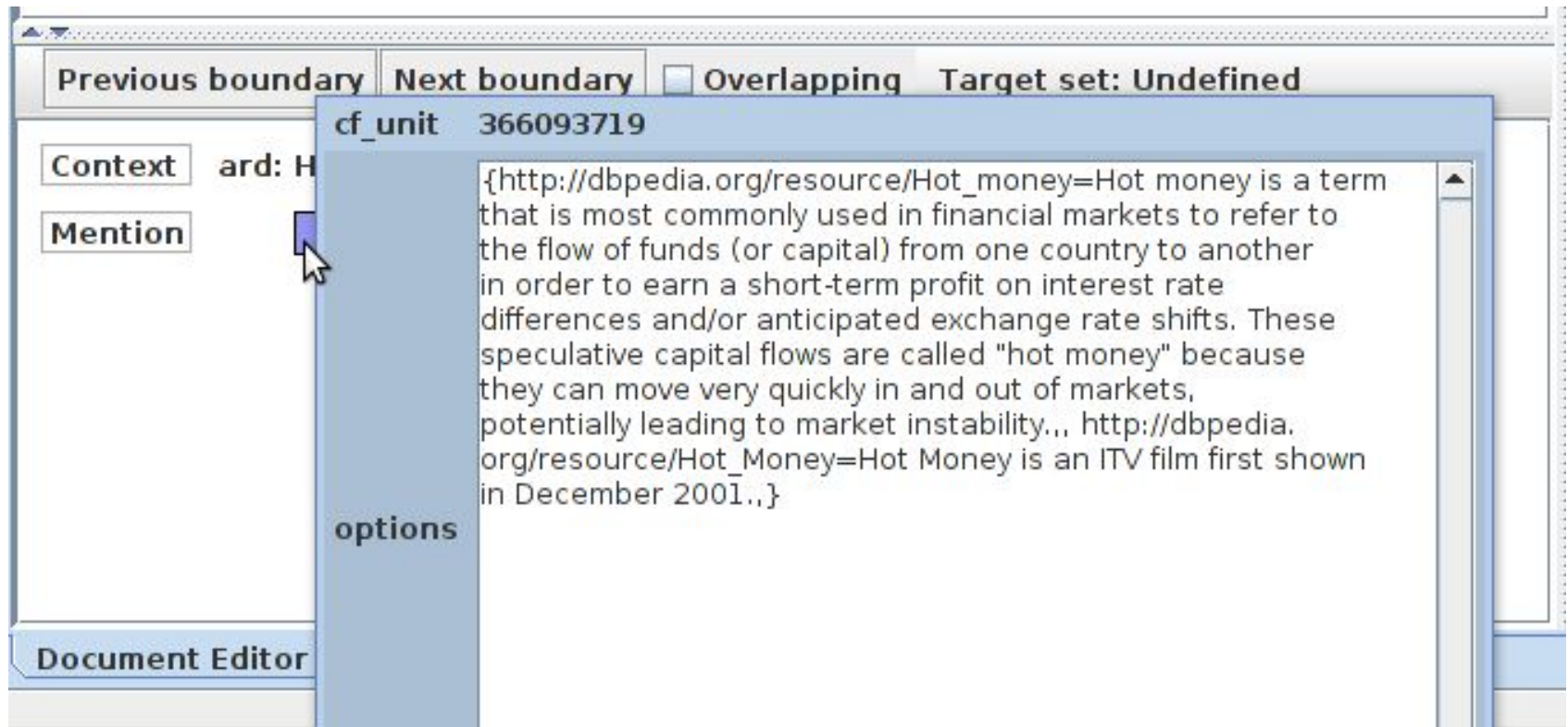
- The task here is to disambiguate named entities by assigning them DBpedia URIs (values of the inst feature)
- From hands-on-crowdsourcing.zip, unpack the adjudication-exercise directory
- Create a corpus and populate it from that directory (11 docs)
- Double click several of them and try adjudicating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



Automatic Adjudication: Hands on

- Use the same corpus as before (from adjudication-exercise directory) but reload it to remove manual modifications.
- Create a new Majority-vote consensus builder PR (classification) and add to a pipeline.
- Set the minimum agreement to 2 meaning both annotators must agree – keep all other parameters the same.
- Run the pipeline and check the crowdConsensus and crowdDisputed sets.

Hands on: Questions



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



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 Computational 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. Sindhvani. 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.