Crowdsourcing Named Entity Recognition and Entity Linking Corpora

Kalina Bontcheva, Leon Derczynski, Ian Roberts

Abstract This chapter describes our experience with crowdsourcing a corpus containing named entity annotations and their linking to DBpedia. The corpus consists of around 10,000 tweets and is still growing, as new social media content is added. We first define the methodological framework for crowdsourcing entity annotated corpora, which combines expert-based and paid-for crowdsourcing. In addition, the infrastructural support and reusable components of the GATE Crowdsourcing plugin are presented. Next, the process of crowdsourcing named entity annotations and their DBpedia grounding is discussed in detail, including annotation schemas, annotation interfaces, and inter-annotator agreement. Where different judgements needed adjudication, we mostly used experts for this task, in order to ensure a high quality gold standard.

1 Introduction

Research on information extraction, and named entity recognition in particular, has been driven forward by the availability of a substantial number of suitably annotated corpora, e.g. ACE [1], MUC [10], and CoNLL [37].[1] These corpora underpin algorithm training and evaluation and predominently contain longer, newspaper-style documents. Unfortunately, when Named Entity Recognition (NER) algorithms are trained on such texts, they tend to perform poorly on shorter, noisier, and more colloquial social media content, as noted by [32, 14].

The problem stems from the very limited amount of social media gold standard datasets currently available. In particular, prior to the start of our project, there were

All authors
University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, UK, e-mail: Initial.Surname@dcs.shef.ac.uk

[1] http://www.clips.uantwerpen.be/conll2003/ner/

fewer than 10,000 tweets annotated with named entities [15, 32].[2] As part of the uComp research project,[3] we aimed to address this gap by creating two kinds of microblog corpora: one annotated with named entities (i.e. names of persons, locations, organisations, and products) and a second, entity linking one, where entity mentions are disambiguated and linked to an external resource [31].

However, creating new sufficiently large datasets through traditional expert-based text annotation methods alone is very expensive, both in terms of time and funding required. The latter has been shown to vary between USD 0.36 and 1.0 per token for some semantic annotation tasks, though NER was not investigated [28], which is unaffordable for smaller-scale research projects like ours. Even though some cost reductions can be achieved through web-based collaborative annotation tools, such as GATE Teamware [7], these can still be costly.

Instead, we experimented with commercial crowdsourcing marketplaces, which have been reported to be 33% less expensive than in-house employees on tasks such as tagging and classification [16]. In order to ensure quality, our corpus annotation approach is based on combining high quality expert-sourcing of annotations, with the scale and quick turn around offered by the paid-for CrowdFlower marketplace [4].[4]

The rest of this use case chapter is structured as follows. Section 2 defines the methodological framework for crowdsourcing annotated corpora, which we adopted. Next, Section 3 describes our crowdsourced named entity recognition dataset, followed by the entity linking corpus (Section 4), conclusions, and future work. It must be noted that a comprehensive survey of crowdsourcing is beyond the scope of this chapter, however, for details see Chapter **??** for guidelines for crowdsourcing in corpus collection, see [34].

2 Corpus Annotation Methodology

Conceptually, the process of crowdsourcing corpora can be broken down into a set of steps (see Figure 1), which form a common methodological framework for the two crowdsourcing case studies (named entity annotation and entity linking respectively).

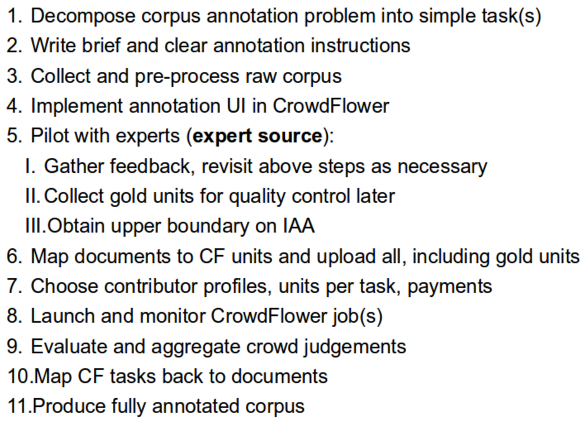
The mapping between an NLP annotation task and a suitable CrowdFlower task workflow is project specific and will thus be discussed in the respective sections. The same applies to the instructions which will be shown to the crowdworkers. Data collection is discussed in Section 2.1, whereas pre-processing is also specific to each NLP annotation task.

The CrowdFlower User Interfaces (UIs) tend to fall into a set of categories, the most commonly used being selection, categorisation, and text input. These can be

[2] A corpus of 12 245 tweets with entity annotations was created by [24], but this is not shared due to Microsoft policy and the system is not available either.

[3] http://www.ucomp.eu

[4] http://www.crowdflower.com



1. Decompose corpus annotation problem into simple task(s)
2. Write brief and clear annotation instructions
3. Collect and pre-process raw corpus
4. Implement annotation UI in CrowdFlower
5. Pilot with experts (**expert source**):
   I. Gather feedback, revisit above steps as necessary
   II. Collect gold units for quality control later
   III. Obtain upper boundary on IAA
6. Map documents to CF units and upload all, including gold units
7. Choose contributor profiles, units per task, payments
8. Launch and monitor CrowdFlower job(s)
9. Evaluate and aggregate crowd judgements
10. Map CF tasks back to documents
11. Produce fully annotated corpus

Fig. 1 The methodological framework for crowdsourcing named entity annotations

generalised and reused between annotation projects, which motivated us to provide reusable, open-source implementations as part of the new GATE Crowdsourcing plugin (see Section 2.5 for details).

The next step is *the expert sourcing pilots*. A sample of the data, the instructions, and the UI are launched in CrowdFlower, exactly as they would appear to the paid-for workers. NLP researchers and/or domain experts were emailed the pilot URL and asked to complete the annotation micro tasks in CrowdFlower and provide detailed feedback (via email). In this way, firstly, the entire annotation workflow is tested *in vitro* and changes to task design, instructions, and CrowdFlower UIs are made as required. Secondly, once all above become stable, expert annotated data is gathered and later used by CrowdFlower as test units, for automatic quality control. In our experience, around 10% of gold units need to be expert sourced, in order to ensure high quality crowd annotations later. Thirdly, the inter-expert agreement gives us a useful estimate of what is the highest achievable agreement between crowd workers.

Once the expert-sourcing pilot is complete, the raw corpus needs to be mapped to CrowdFlower units and uploaded in the system (step 6). In earlier prototypes we used Python scripts to map between documents in GATE [12] format and CrowdFlower units, then bulk uploaded the data in a spreadsheet format. Now, however, we make use of the GATE Crowdsourcing plugin [9], which not only does the mapping and upload automatically, but also later imports the crowd judgements back into the GATE corpus and documents (steps 10 and 11).

Step 7, choosing contributor profiles, units per task and payments, is essentially the configuration of the crowdsourcing project's execution. CrowdFlower enables us to restrict crowdworkers based on their country and past performance, as well as define the maximum number of units that individual crowd workers are allowed to complete. This is particularly important, since typically there is a group of highly active contributors, who could otherwise introduce significant annotator bias into the corpus.

CrowdFlower also provides control over the number of annotation units per page shown to the crowdworkers (i.e. task). This is another important parameter, which impacts annotation quality. In particular, when expert-sourced gold units are provided to CrowdFlower, it will automatically mix a gold unit amongst the unannotated units. If an annotator performs poorly on the gold unit, the entire task will be discarded, thus reducing spam and improving overall quality. For both use cases, we used 5 units per task. This, in our experience, provided a good balance between annotation quality and task size.

Lastly, with respect to payments, we set 6 cents per entity annotation task and 6 cents per entity linking task.[5] CrowdFlower automatically carries out contributor satisfaction surveys for each worker, which allowed us to fine tune pay, instructions, and ease of the test/gold units. For this purpose, we ran small paid-for pilots on 500 unannotated units each.

Once all these parameters are set, the CrowdFlower project can be launched and monitored through the CrowdFlower web consoles (step 8). There are also facilities for monitoring quality and inter-worker agreement per unit (step 9).

The rest of this section provides more details on how we collected the raw tweet corpus; how annotations were represented; quality control; corpus production/use; and the GATE Crowdsourcing plugin.

2.1 Data collection

The main source of data was one of Twitter's public streaming feeds. In this case, we used the garden hose feed, which provides a random 10% sample of all tweets. Within this feed, we geolocated 250,000 users as inside the UK using a graph based system [33] and captured their public activity for a number of months. The result was a large collection of JSON representing tweets from accounts based in the UK. These were then tagged with language information [29]. This corpus was also used to survey voter intent [21].

	Total	**Without NE mention**	**With NE mention**	**% with NEs**
Overall	2394	1474	920	38.4
Non-directed	1803	1012	791	43.9
Directed	591	462	129	21.8

Table 1 Distribution of named entities in directed and non-directed tweets.

Informal examination of the tweets suggested that a large number of them were directed, i.e. began with a user mention and were not displayed in a user's subscrip-

[5] The resulting median pay for trusted contributors on entity recognition was USD$11.37/hr, an ethical rate of pay considering that the majority of crowdsource workers rely on it for income.

tion stream but rather directed at a specific user or users. Working on the premise that this might have an impact on the likelihood of named entities being present, we surveyed the distribution of entity mentions between directed and non-directed tweets in an existing gold-standard corpus [32]. Results are shown in Table 2.1. We found that non-directed tweets were more likely to bear named entities, and so discarded any directed tweets. This dataset was then shuffled and the 10,000 chosen for further analysis.

An important part of data collection in the case of social media content, in particular, is screening for offensive content. This is neccessary to access a greater crowd and not just those workers who have opted-in to work with adult material. It does introduce a potential bias in corpus construction and composition, but our focus here is on crowdsourcing, and the proportion of tweets involved is minor. We use the BBC objectionable terms list[6] and remove tweets with any matching words. Further, we remove objectionable tweets noticed by humans in the expert-sourcing stage (our annotators returned some objectionable tweets in this closed stage of the annotation process). Finally, we monitor Twitter slang dictionaries and use these to build lists of shortened adult/offensive terms; tweets containing these are also removed.

2.2 Physical representation

Documents and their annotations are encoded in the GATE stand-off XML format [12], which was chosen for its support for overlapping annotations and the wide range of automatic pre-processing tools available. GATE also has support for the XCES standard [17], if this is preferred. Annotations are grouped in separate annotation sets: one for the automatically pre-annotated annotations, one for the crowdsourced judgements, and a consensus set, which can be considered as the final gold standard annotation layer. In this way, full provenance is tracked and also, it is possible to experiment with methods, which consider more than one answer as potentially correct.

2.3 Quality control

The key mechanism for spam prevention and quality control in CrowdFlower is through test units, which we also refer to as gold units. We recruited a pool of 14 NLP expert volunteers and, using CrowdFlower, piloted the annotation project, as discussed above. In this way a set of 2,000 tweets were annotated with gold-standard named entities. A subset (500) of these were then also entity disambiguated by the volunteers. The latter gold named entity linking (NEL) set is in the process of being

[6] Prefaced at http://www.bbc.co.uk/editorialguidelines/page/guidance-language-full

expanded further, but even its smaller size proved sufficient for generating gold units for the paid-for jobs.

For the adjudication step, both in the case of the expert-sourced and the crowd-sourced data, we asked the best performing NLP experts to use GATE's annotation stack editor and reconcile any remaining differences. As a first step, we used JAPE rules [11] to identify automatically all cases where volunteers/crowdworkers disagreed and these were then adjudicated manually. For the second phase of our corpus crowdsourcing, we are planning on feeding these annotations back into a new CrowdFlower project automatically. This can then either be expert-adjudicated as before or verified through gathering additional crowd judgements.

2.4 Usage

Distributing microblog corpora to other researchers is a difficult task, due to Twitter's term of service. As part of GATE, we are working on a corpus export function, which provides a list of tweet identifiers and encodes the linguistic annotations in a JSON datastructure. In this way, each researcher will be able to download the tweets afresh and then merge them with our annotations, to obtain the complete corpus. When completed, the corpus will be made available for download from http://gate.ac.uk, but in the mean time, the data is available upon request.

In our experience, this approach, albeit legally necessary, is far from ideal, since tweets can be deleted in the mean time and/or URLs contained within can stop being accessible. This impacts replicability and can make it hard to run comparative evaluations, especially if the social media content is more than 1 year old.

2.5 Infrastructural Support

As can be seen from Figure 2.1, each corpus crowdsourcing project has to address a number of largely infrastructural steps. In particular, corpus pre-annotation and transformation into CrowdFlower units, creation of the annotation UI, creation and upload of the gold units, and finally mapping units back into documents and aggregating all judgements to produce the gold standard.

Based on our corpus annotation experience, we implemented a generic, open-source GATE Crowdsourcing plugin, which makes it very easy to setup and carry out crowdsourcing-based corpus annotation from within GATE's visual interface. The plugin contains reusable task definitions and crowdsourcing user interface templates which can be used by researchers to commission CrowdFlower jobs directly from within GATE's graphical user interface. They can also pre-process the data automatically with relevant GATE linguistic analysers, prior to crowdsourcing. Once all parameters are configured, the new crowdsourcing job builder generates the respective CrowdFlower units automatically.

The CrowdFlower web interface is still necessary, in order to configure the contributors and launch the project.

On completion, the collected multiple judgements are imported automatically back into GATE and the original documents are enriched with the crowdsourced information, modelled as multiple annotations (one per contributor). GATE's existing tools for calculating inter-annotator agreement and corpus analysis can then be used to gain further insights into the quality of the collected information.

3 Crowdsourcing a Named Entity Annotated Corpus

Named entity annotation is often regarded as a sequential labelling problem [23], where the crowd workers select a contiguous text chunk and then choose its entity category. This NE annotation task design is closer to the way expert-based annotation tools for NEs operate (e.g. GATE Teamware [7]). Yet, in a crowdsourcing context, sequence labeling is difficult to implement and could raise issues with annotation recall if the incentives mechanisms are not suitably designed. Concretely, two issues were identified when using Amazon Mecanical Turk (MTurk) for this task [23]:

- the MTurk interface did not allow text selection and
- the per-document, fixed payment rate encourages annotators to only mark entities in the first few sentences thus leading to low recall.

Lawson et al. [23] addressed this by building a custom user interface and providing a new incentive model where annotators are payed a fixed rate for each document and then gain additional bonuses for each additional NE that they identify. Their interface allows workers to identify text boundaries and to label those with an NE category.

A classification-based crowdsourcing approach has also been experimented with [15]. In this task design, annotators are shown one sentence (or tweet) per unit and have to mark each word in the sentence as belonging to one of a given set of entity types (e.g. Person, Location, Organisation) or no entity [15, 22]. The downside of this approach is that it is hard to fit more than one unit per CrowdFlower task, since there needs to be space for the words by NE types grid of check boxes. As sentence length grows, so does the screen size required per unit. In addition, this grid-like design with check boxes is not very fast to annotate with, since users need to make a selection for each word in the sentence (choosing between the 3-4 entity types or None).

Further schemata are available for annotation, especially in the case where multiple classes may potentially be marked over the same sequences when only one may apply. For example, annotators may first select entity mention bounds and then perform entity classification as a second step. Such schemata are discussed in [38].

In the rest of this section, we discuss in detail an alternative NE annotation task design, the corresponding user interface, and the corpus crowdsourcing process.

3.1 NE Task Design and Data Preparation

The single most influential part of any linguistic annotation exercise is the annotator's ability to clearly understand and conduct the annotation task. This is controlled to some degree by both the annotation guidelines and the annotation tool. Having simple, short guidelines that include examples and specific instructions is helpful. Further, having a clean interface is important – more important that having an interface in one's native language [19] – and as with all web design, interaction should be simple and intuitive [20].

In order to overcome the problems of the previous NE task designs, we developed an approach, which aims to combine the user interface compactness of the sequential labelling design with the nicely constraining nature of classification approaches.

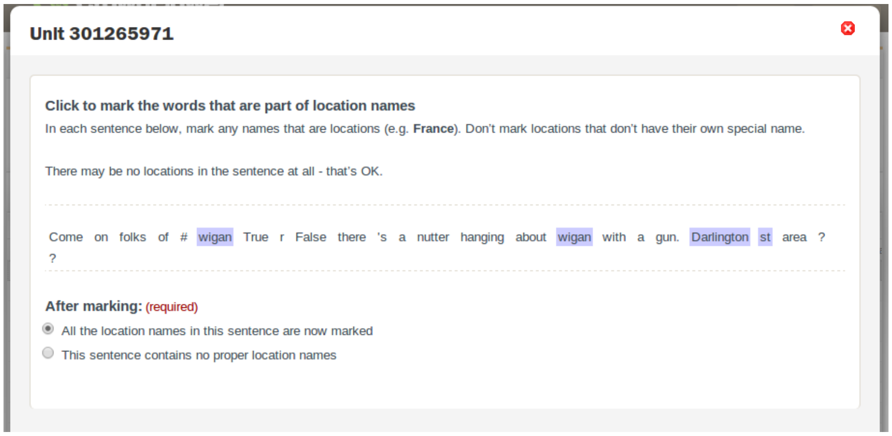
Firstly, documents are pre-segmented into sentences and word tokens, using GATE's TwitIE plugin [8], which provides a tokeniser, POS tagger, and a sentence splitter, specifically adapted to microblog content. Due to the short length of tweets, we also opted to show one tweet per CrowdFlower unit. The GATE Crowdsourcing plugin can be configured easily for different mappings, e.g. to show one sentence per unit or one paragraph. In our experience, for NE annotation a sentence or a tweet provide sufficient context.

Next, each tweet and the words contained within are loaded into CrowdFlower and the user interface configured so that contributors are asked to click only on words which constitute an entity, using custom JavaScript and CSS inserted into the CrowdFlower form (see Figure 2). For multi-word entities users are expected to click on all words separately.

Lastly, in order to keep our task descriptions very short, while also giving as many positive and negative examples as possible, we decided to generate a separate CrowdFlower job for each named entity type. In particular, we focused on annotating persons, organisations, locations, and products. Thus four projects were generated and each tweet was inserted for annotation within each of these four projects. The benefit from annotating each entity type separately is that contributors are primed to focus on one kind of entity only, as they go through the CrowdFlower tasks. The drawback is obviously higher costs, since each tweet gets annotated effectively 12 times (3 contributors x 4 named entity types).

3.2 The Named Entity Annotation Interface

We designed a CrowdFlower-based user interface for word-constrained sequential selection (see Figure 2), which we hope is easily reusable for other similar annotation tasks. The annotators are instructed to click on all words that constitute an entity of the particular type (e.g. location). Adjacent entities of the same type end up being run together, as with the CoNLL representation, which is not preferred, but a low-impact side effect of the simple user interface – and a tolerable cost of having an easy, friendly task environment.



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

Fig. 2 The Named Entity Selection Interface in CrowdFlower

Since tweets may not contain any entity of the target type, we have also added an explicit confirmation step. This forces annotators to declare that they have selected all entities or that there are no entities in this tweet. In this way, CrowdFlower can then use gold units and test the correctness of the selections, even in cases where no entities are present in the tweet.

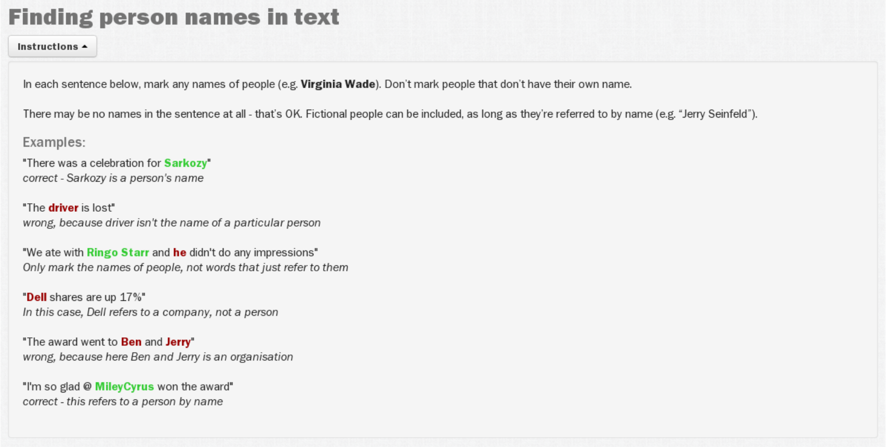
3.3 The NE Corpus Annotation Process

3.3.1 The NE Expert-Sourcing Pilot

Expert-sourcing has two distinct benefits. Firstly, it provides a verbose channel for feedback on task design which is often unused or even impossible for wider crowdsourcing tasks. Secondly, the resulting "gold" data is better: it is high-quality (as it has been checked by multiple annotators) and it is broad (as there tends to be a reasonably large amount of resulting data) covering a wider range of situations and potential edge cases with which to guide unskilled crowdsourced worked.

We expert-sourced annotations of named entities over 3000 non-directed tweets (i.e. those not beginning with an @username, since our experience from previous datasets is that these are more likely to contain NEs). Four types of entity were gathered per tweet (product, person, location, organisation), and each tweet was annotated by two expert volunteers, using our user interface.

Since crowd tasks were given for one entity type at a time, there were a total of 12,000 sets of annotations made (3000 tweets * 4 entity types * 2 annotations). An expert adjudicator then manually checked the results to create a consensus set. Note that because of the complex nature of the annotation task and the increased variation in actual annotator inherent in expert- and crowdsourcing, we have not yet



In each sentence below, mark any names of people (e.g. **Virginia Wade**). Don't mark people that don't have their own name.

There may be no names in the sentence at all - that's OK. Fictional people can be included, as long as they're referred to by name (e.g. "Jerry Seinfeld").

Examples:

"There was a celebration for **Sarkozy**"
correct - Sarkozy is a person's name

"The **driver** is lost"
wrong, because driver isn't the name of a particular person

"We ate with **Ringo Starr** and **he** didn't do any impressions"
Only mark the names of people, not words that just refer to them

"**Dell** shares are up 17%"
In this case, Dell refers to a company, not a person

"The award went to **Ben** and **Jerry**"
wrong, because here Ben and Jerry is an organisation

"I'm so glad @ **MileyCyrus** won the award"
correct - this refers to a person by name

Fig. 3 Instructions for person entity recogntion, after feedback created during expert sourcing.

calculated an overall Inter-Annotator Agreement. This may be achieved using more advanced metrics; see [3, 30].

For named entity recognition, questions were generally raised around entity classes and also what types of entity were valid. We refactored these into task instructions and design during expert sourcing. An example set of instructions for the person class is shown in Figure 3.

3.3.2 The NER Crowdsourcing Runs

Having amassed a reasonable amount of expert sourced gold data, we took two runs for named entity recognition annotation; one for each of person and product annotations. The task was to mark the tokens in a given tweet that were names of people or names of products. We used 100 expert sourced gold tweet annotations and 475 UK-based tweets filtered for objectionable content to be annotated by the crowd. We configured two annotators per unit and six units per task. Workers could be from any English-speaking nation. The project was launch mid-afternoon in the UK / morning in the USA. The jobs were completed in less than an hour, gathering 1900 judgments in total.

In our runs, we included the maximum amount of gold data possible (33% of all units), using the expert sourced data for this. At least one unit is shown per task, and workers must qualify by getting a high score on some tasks made up entirely of gold data before progressing to annotate previously-unseen data. In our case, 76% of workers passed this quiz phase without giving up. This 76% then achieved a 97% overall agreement with the expert sourced examples inserted into their annotation tasks. On the data that the crowdsourced workers were annotating, agreement over which words were entities was 98.81% at the token level – although this includes a large number of "not-an-entity" annotations. Note that if all workers annotated nothing, a baseline measure, they would achieve 100% agreement, and so this metric remains questionable in a crowd scenario.

As ongoing work, we are running larger paid-for projects, with the ultimate goal to collect circa 10,000 named-entity expert sourced annotated tweets, in batches of 500 and 1,000. This gives us scope to experiment with expert sourcing scenarios and the scope to cheaply collect an open, large twitter corpus.

4 Crowdsourcing an Entity Linking Corpus

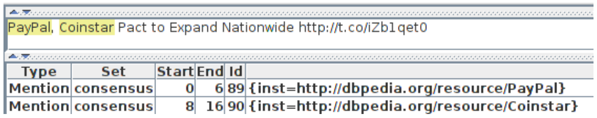
Having determined which expressions in text are mentions of entities, a follow-up task is entity linking. It entails annotating a potentially ambiguous entity mention (e.g. Paris) with a link to a canonical identifier describing a unique entity (e.g. http://dbpedia.org/resource/Paris). Different entity databases have been used as a disambiguation target (e.g. Wikipedia pages in TAC KBP [18]) and Linked Open Data resources (e.g. DBpedia [26], YAGO [35]).

Microblog named entity linking (NEL) is a relatively new, underexplored task. Research in this area has focused primarily on *whole-tweet entity linking* (e.g. [2, 25]), also referred to as an "aboutness" task. The whole-tweet NEL task is defined as determining which topics and entities best capture the meaning of a microtext. However, given the shortness of microtext, correct semantic interpretation is often reliant on subtle contextual clues, and needs to be combined with human knowledge. For example, a tweet mentioning iPad makes Apple a relevant entity, because there is the implicit connection between the two. Consequently, entities relevant to a tweet may only be referred to implicitly, without a mention in the tweet's text. From a corpus annotation perspective, the aboutness task involves identifying relevant entities at whole-document level, skipping the common NER step of determining entity bounds. Both these variants are particularly difficult in microblog genre text (e.g. tweets) [13].

Our focus however is on *word level entity linking*, where the task is to disambiguate only the named entities which are mentioned explicitly in the tweet text, by assigning an entity identifier to each named entity mention. However, unlike TAC KBP [18] where only one entity mention per document is disambiguated, we annotate all entity mentions with disambiguated URIs (Unique Reference Identifiers).

4.1 NEL Annotation Scheme

Our entity linking annotations are encoded as Mentions, with a start and end offset and an inst feature whose value is a DBpedia URI (see Figure 4). They are currently kept separate from the named entity annotations, but the two annotation layers are co-extensive and can easily be merged automatically, e.g. by using JAPE rules (see Section XXX in Chapter YYY **??**).



PayPal, Coinstar Pact to Expand Nationwide http://t.co/iZb1qet0

Type	Set	Start	End	Id	
Mention	consensus	0	6	89	{inst=http://dbpedia.org/resource/PayPal}
Mention	consensus	8	16	90	{inst=http://dbpedia.org/resource/Coinstar}

Fig. 4 Word Level Entity Linking Annotations, shown in GATE

We chose DBpedia [5] as the target entity linking database, due to its good coverage of named entities, its frequent updates, and available mappings to other Linked Open Data resources, such as YAGO [36] and Freebase [6].

4.2 Task Design and Data Preparation

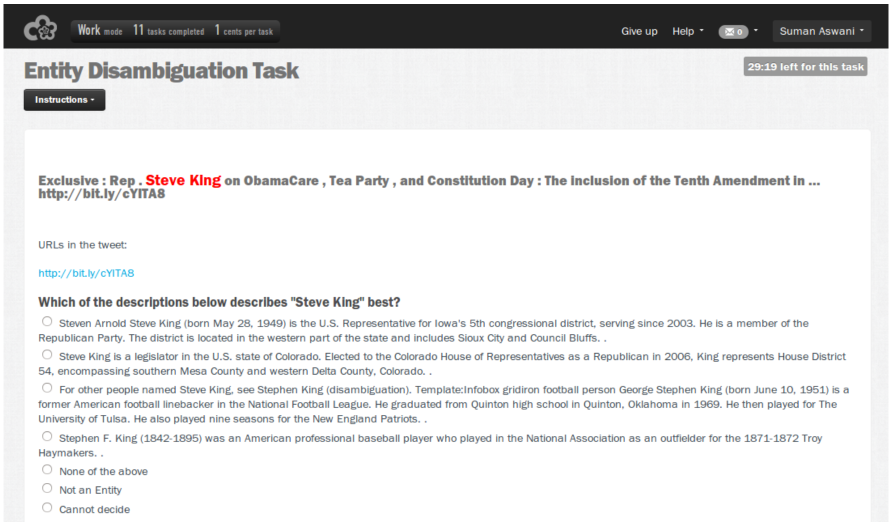
NEL is essentially a classification task, where the goal is to choose amongst one of the possible entity targets from the knowledge base or NIL (no target entity), in cases where no such entity exists. The latter case is quite common in tweets, where people often refer to friends and family, for example. An added problem, however, is that highly ambiguous entity mentions (e.g. Paris), could have tens or even over a hundred possible target entities. Since showing so many options to a human is not feasible, instead, during data preparation, candidate entity URIs are ranked according to their Wikipedia commonness score [27] and only the top 8 are retained and shown, in addition to NIL (which we called "none of the above") and "not an entity" (to allow for errors in the automatic pre-processing). We chose to show at most 8 entity candidates, following a small-scale pilot with NLP experts, which gave us feedback.

In order to overcome the problem that the correct candidate entity could have been present in DBpedia, but filtered out due to low occurrence in Wikipedia, we are implementing an iterative step only for entities where annotators have chosen "none of the above". In those cases, if more than 8 candidates were present originally, we then take the next 8 candidate URIs and repeat the process.

An alternative approach would be to allow NLP experts, who are also familiar with DBpedia, to search and identify the correct entity URI manually. We did not have enough such volunteers to try this.

As can be seen above, the key data preparation step is the generation of candidate entity URIs. Even though error prone, candidate entity selection against DBpedia needs to be carried out automatically, since the latest English DBpedia contains 832,000 persons, 639,000 places, and 209,000 organisations, out of the 4 million DBpedia URIs in total.

Relying purely on looking up exact matching labels in DBpedia is not sufficient, since entities are often referred to by acronyms, nicknames, and shortened names (e.g. surnames like Obama or Snowden). Instead, we match the string of the named entity mention in the document (annotated already in the corpus) against the values of the *rdf:label*, *foaf:name* and several other similar annotation properties,


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/cYITA8

URLs in the tweet:

http://bit.ly/cYITA8

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

Fig. 5 The Named Entity Linking (NEL) Interface in CrowdFlower

for all instances of the *dbpedia-ont:Person*, *dbpedia-ont:Organisation* and *dbpedia-ont:Place* classes in DBpedia. Acronyms and shorter ways of referring to DBpedia entity URIs are collected also from the Wikipedia anchor texts, that point to the respective Wikipedia page.[7]

Lastly, we had to choose the size of the context, shown to the annotators to help with the disambiguation of the entity mention. We experimented with showing the sentence where the mention appears, but this was not sufficient. Therefore, we show the entire tweet text and any web links within. For longer documents, it would make sense to show at least 1 preceding and 1 following sentence, or even the containing paragraph, space permitting.

4.3 The NEL Annotation Interface

We designed a CrowdFlower-based user interface (see Figure 5), which showed the text of the tweet, any URL links contained therein, and a set of candidate targets from DBpedia. The instructions encouraged the annotators to click on the URL links from the tweet, in order to gain addition context and thus ensure that the correct DBpedia URI is chosen.

Candidate entity meanings were shown in random order, using the text from the corresponding DBpedia abstracts (where available) or the actual DBpedia URI otherwise.

[7] There is a 1-to-1 mapping between each DBpedia URI and the corresponding Wikipedia page, which makes it possible to treat Wikipedia as a large corpus, human annotated with DBpedia URIs.

In addition, the options “none of the above” and “not an entity” were added, to allow the annotators to indicate that this entity mention has no corresponding DBpedia URI (none of the above) or that the highlighted text is not an entity. In the expert-only version, we added a third option “cannot decide”, so the volunteers could indicate that the context did not provide sufficient information to reliably disambiguate the entity mention. However, this was removed in the paid-for crowdsourcing version, in order to discourage the crowd workers from always choosing this option as the quick and easy choice.

4.4 NEL Annotation Process

The corpus annotation process involved two main stages: the expert-sourcing pilot stage and the paid-for crowdsourcing stage. We describe each of them in more detail next. With respect to annotator instructions, we only showed a short paragraph stating:

```
This task is about selecting the meaning of a word or a phrase.
You will be presented with a snippet of text within which one
or more words will be highlighted. Your task is to select the
option, which matches best the meaning of the highlighted text.

If you are sure that none of the available options are correct
then select None of the above.

If the highlighted text is not a name of something, then select
 Not an entity.
```

4.4.1 NEL Expert-Sourcing Pilot

We chose a random set of 177 entity mentions for the expert-sourcing NEL pilot and generated candidates URIs for them. Each entity mention was disambiguated by a random set of three NLP experts, using our NEL annotation interface. We had 10 volunteer experts in total for this pilot.

Annotations for which no clear decision was made were adjudicated by a fourth expert who had not previously seen the tweets.

As each entity annotation was disambiguated by three NLP volunteers, we determined agreement by measuring the proportion of annotations on which all three made the same choice. Out of the resulting 531 judgements, unanimous inter-annotator agreement occurred for 89% of entities. The resulting expert-sourced dataset consisted of 172 microblog texts, containing entity mentions and their assigned DBpedia URIs. This was used as gold units for the paid-for crowdsourcing annotation.

4.4.2 NEL Crowdsourcing Runs

Next, 400 tweets containing 577 entity mentions were loaded as CrowdFlower units, using the GATE Crowdsourcing plugin. Automatic candidate generation of DBpedia URIs was carried out, as described above. The 177 gold units were loaded into the system, for use as test questions. We configured three annotators per unit and five units per task, one of which is from our test set. CrowdFlower does the unit mixing and test question selection automatically. A payment of 6 cents per task was offered.

We restricted contributors to those from English speaking countries only and launched the project earlier in the morning in the UK. In this way, we hoped that crowdworkers from the UK could start on the jobs first. This decision was motivated by the fact that our tweets were collected from UK-based tweeters and were discussing many UK-specific entities (e.g. local football teams, UK niche artists).

The job was completed in less than 4 hours, by 11 contributors (5 from the UK, 3 from Canada and 3 from the US). The CrowdFlower reported agreement of 67% between them. All contributors had passed the test units 100% successfully and were not spammers. The disagreements came from flaws in our automatic candidate selection pre-processing, where in addition to creating candidates for entities like Paris, candidates were also created for all words tagged as NNP or NNPs (i.e. proper names) by our part-of-speech tagger. This unfortunately, resulted in spurious candidates being generated for words like “Happy” and “Mum”. Some contributors would in that case choose “not an entity”, whereas others – “none of the above”.

In subsequent crowdsourcing runs, we therefore ensured that automatically generated candidates are better aligned with the crowdsourced named entities. This raised agreement to on average 80%, which compares favourably to the 89% agreement achieved by our expert volunteers. We kept the rest of the settings unchanged.

Adjudication on the first 1,000 crowdsourced NEL tweets was carried out by two of our NLP expert volunteers, who had not seen the tweets previously. They were presented with the choices made by the CrowdFlower contributors and asked to choose amongst them. This is an easier and faster to carry out, rather than showing all available candidate URIs.

We are in the process of crowdsourcing another 3,000 to 5,000 NEL tweets, where we will adjudicate by soliciting additional judgements only on the contentious cases and taking a majority vote.

5 Conclusion

This chapter presented two related case studies in crowdsourcing annotations for training and evaluation of named entity recognition and entity linking algorithms. The focus was specifically on acquiring high-quality annotations of social media content. This motivated our methodology, which combines expert-sourcing and paid-for crowdsourcing, in order to ensure quality at affordable costs. Where different judgements needed adjudication, we mostly used experts for this task.

In future work we plan on extending the crowdsourced corpora further and at the same time, to continue improving the new GATE Crowdsourcing plugin, which reduces significantly the overhead of mapping NLP corpus annotation tasks onto CrowdFlower units and then back, aggregating all these judgements to produce the final corpus. In particular, we plan to implement automatic IAA calculation for the multi-project sequential selection tasks which arise in the case of named entity annotation. We are also in the process of implementing a JSON-based corpus exporter, which will allow us to distribute freely the collected annotations, coupled only with tweet IDs.

Another line of research, as part of the uComp project, will be on experimenting with games with a purpose, instead of and in addition to paid-for crowdsourcing.

Lastly, based on the newly collected corpora, we are in the process of training and evaluating different machine learning algorithms for named entity recognition and entity linking, specifically on short, microblog content.

Acknowledgements Special thanks to Niraj Aswani for implementing the initial entity linking prototype in CrowdFlower, as well as to Marta Sabou, Arno Scharl, and other uComp project members for the feedback on the task and user interface designs. Also, many thanks to Johann Petrak and Genevieve Gorrell for their help with the automatic candidate generation for entity linking. We are particularly grateful to all researchers at the Sheffield NLP group and members of the TrendMiner and uComp projects, who helped create the gold data units. This research has received funding support of EPSRC EP/K017896/1, FWF 1097-N23, and ANR-12-CHRI-0003-03, in the framework of the CHIST-ERA ERA-NET (uComp project), as well as the UK Engineering and Physical Sciences Research Council (grant EP/I004327/1).

References

1. ACE: Annotation Guidelines for Event Detection and Characterization (EDC) (Feb 2004), available at http://www.ldc.upenn.edu/Projects/ACE/
2. Amigó, E., Corujo, A., Gonzalo, J., Meij, E., Rijke, M.d.: Overview of RepLab 2012: Evaluating Online Reputation Management Systems. In: CLEF 2012 Labs and Workshop Notebook Papers (2012)
3. Artstein, R., Poesio, M.: Kappa3 = Alpha (or Beta). Tech. Rep. CS Technical Report CSM-437, Department of Computer Science, University of Essex, Colchester, UK (2005)
4. Biewald, L.: Massive multiplayer human computation for fun, money, and survival. In: Current Trends in Web Engineering, pp. 171–176. Springer (2012)
5. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia – a crystallization point for the web of data. Journal of Web Semantics: Science, Services and Agents on the World Wide Web 7, 154–165 (2009)
6. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data. pp. 1247–1250. ACM (2008)
7. Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Aswani, N., Gorrell, G.: GATE Teamware: A Web-based, Collaborative Text Annotation Framework. Language Resources and Evaluation 47, 1007—1029 (2013)
8. Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M.A., Maynard, D., Aswani, N.: TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In: Proceed-

ings of the International Conference on Recent Advances in Natural Language Processing. Association for Computational Linguistics (2013)

9. Bontcheva, K., Roberts, I., Derczynski, L., Rout, D.: The GATE Crowdsourcing Plugin: Crowdsourcing Annotated Corpora Made Easy. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Association for Computational Linguistics (2014)
10. Chinchor, N.A.: Overview of MUC-7/MET-2. In: Proceedings of the 7^{th} Message Understanding Conference (MUC7) (Apr 1998), available at http://www.muc.saic.com/proceedings/muc_7_toc.html
11. Cunningham, H.: JAPE: a Java Annotation Patterns Engine. Research Memorandum CS–99–06, Department of Computer Science, University of Sheffield (May 1999)
12. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: an Architecture for Development of Robust HLT Applications. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 7–12 July 2002. pp. 168–175. ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002), http://gate.ac.uk/sale/acl02/acl-main.pdf
13. Derczynski, L., Maynard, D., Aswani, N., Bontcheva, K.: Microblog-Genre Noise and Impact on Semantic Annotation Accuracy. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media. ACM (2013)
14. Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Bontcheva, K.: Analysis of named entity recognition and linking for tweets. Information Processing and Management 51, 32–49 (2015)
15. Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., Dredze, M.: Annotating named entities in Twitter data with crowdsourcing. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. pp. 80–88 (2010)
16. Hoffmann, L.: Crowd control. Communications of the ACM 52(3), 16 –17 (2009)
17. Ide, N., Bonhomme, P., Romary, L.: XCES: An XML-based Standard for Linguistic Corpora. In: Proceedings of the second International Conference on Language Resources and Evaluation (LREC 2000), 30 May – 2 Jun 2000. pp. 825–830. Athens, Greece (2000), http://www.lrec-conf.org/proceedings/lrec2000/pdf/172.pdf
18. Ji, H., Grishman, R., Dang, H.T., Griffitt, K., Ellis, J.: Overview of the tac 2010 knowledge base population track. In: Proceedings of the Third Text Analysis Conference (2010)
19. Khanna, S., Ratan, A., Davis, J., Thies, W.: Evaluating and improving the usability of Mechanical Turk for low-income workers in India. In: Proceedings of the first ACM symposium on computing for development. ACM (2010)
20. Krug, S.: Don't make me think: A common sense approach to web usability. Pearson Education (2009)
21. Lampos, V., Preotiuc-Pietro, D., Cohn, T.: A user-centric model of voting intention from social media. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. pp. 993–1003. Association for Computational Linguistics (2013)
22. Laws, F., Scheible, C., Schütze, H.: Active learning with amazon mechanical turk. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1546–1556 (2011)
23. Lawson, N., Eustice, K., Perkowitz, M., Yetisgen-Yildiz, M.: Annotating large email datasets for named entity recognition with mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. pp. 71–79 (2010)
24. Liu, X., Zhou, M., Wei, F., Fu, Z., Zhou, X.: Joint inference of named entity recognition and normalization for tweets. In: Proceedings of the Association for Computational Linguistics. pp. 526–535 (2012)
25. Meij, E., Weerkamp, W., de Rijke, M.: Adding semantics to microblog posts. In: Proc. of the Fifth Int. Conf. on Web Search and Data Mining (WSDM) (2012)

26. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: DBpedia Spotlight: Shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems (I-Semantics) (2011)
27. Milne, D., Witten, I.H.: Learning to link with Wikipedia. In: Proc. of the 17th Conf. on Information and Knowledge Management (CIKM). pp. 509–518 (2008)
28. Poesio, M., Kruschwitz, U., Chamberlain, J., Robaldo, L., Ducceschi, L.: Phrase Detectives: Utilizing Collective Intelligence for Internet-Scale Language Resource Creation. Transactions on Interactive Intelligent Systems 3 (2013)
29. Preotiuc-Pietro, D., Samangooei, S., Cohn, T., Gibbins, N., Niranjan, M.: Trendminer: An architecture for real time analysis of social media text. In: Proceedings of the workshop on Real-Time Analysis and Mining of Social Streams (2012)
30. Ramanath, R., Choudhury, M., Bali, K., Roy, R.S.: Crowd prefers the middle path: A new iaa metric for crowdsourcing reveals turker biases in query segmentation. In: Proceedings of the annual conference of the Association for Computational Linguistics. vol. 1, pp. 1713–1722 (2013)
31. Rao, D., McNamee, P., Dredze, M.: Entity linking: Finding extracted entities in a knowledge base. In: Multi-source, Multi-lingual Inf. Extraction and Summarization. Springer (2013)
32. Ritter, A., Clark, S., Mausam, Etzioni, O.: Named entity recognition in tweets: An experimental study. In: Proc. of Empirical Methods for Natural Language Processing (EMNLP). Edinburgh, UK (2011)
33. Rout, D., Preotiuc-Pietro, D., Bontcheva, K., Cohn, T.: Wheres @wally? a classification approach to geolocating users based on their social ties. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media (2013)
34. Sabou, M., Bontcheva, K., Derczynski, L., Scharl, A.: Corpus annotation through crowdsourcing: Towards best practice guidelines. In: Proceedings of the 9th international conference on language resources and evaluation (LREC14). pp. 859–866 (2014)
35. Shen, W., Wang, J., Luo, P., Wang, M.: LINDEN: Linking named entities with knowledge base via semantic knowledge. In: Proceedings of the 21st Conference on World Wide Web. pp. 449–458 (2012)
36. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th international conference on World Wide Web. pp. 697–706. ACM (2007)
37. Tjong Kim Sang, E.F., Meulder, F.D.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of CoNLL-2003. pp. 142–147. Edmonton, Canada (2003)
38. Voyer, R., Nygaard, V., Fitzgerald, W., Copperman, H.: A hybrid model for annotating named entity training corpora. In: Proceedings of the fourth linguistic annotation workshop. pp. 243–246. Association for Computational Linguistics (2010)