Undefined 1 (2015) 1–5
IOS Press

Towards hybrid NER: an extended study of content and crowdsourcing-related performance factors

Editor(s): Marta Sabou, Technical University of Vienna, Austria
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Oluwaseyi Feyisetan [a,*], Elena Simperl [b], Markus Luczak-Roesch [b], Ramine Tinati [b], and Nigel Shadbolt [b]
[a] *University of Southampton, Southampton, UK*
E-mail: oof1v13@soton.ac.uk
[b] *University of Southampton, UK*
E-mail: {e.simperl,m.luczak-rosch,r.tinati,nrs}@soton.ac.uk

Abstract. This paper explores the factors that influence the human component in hybrid approaches to named entity recognition (NER) in microblogs, which combine state-of-the-art automatic techniques with human and crowd computing. We identify a set of content and crowdsourcing-related features (number of entities in a post, types of entities, content sentiment, skipped true-positive posts, average time spent to complete the tasks, and interaction with the user interface) and analyse their impact on the accuracy of the results and the timeliness of their delivery. Using CrowdFlower and a simple, custom built gamified NER tool we run experiments on three datasets from related literature and a fourth newly annotated corpus. Our findings show that crowd workers are adept at recognizing people, locations, and implicitly identified entities within shorter microposts. We expect these findings to lead to the design of more advanced NER pipelines, informing the way in which tweets are chosen to be outsourced or processed by automatic tools. Experimental results are published as JSON-LD for further use by the research community.

Keywords: crowdsourcing, human computation, named entity recognition, microposts

1. Introduction

Information extraction is a central component of the Web of Data vision [3]. An important task in this context is the identification of named entities - the people, places, organisations, and dates referred to in text documents - and their mapping to Linked Data URIs [34]. State-of-the-art technology in entity recognition achieves near-human performance for many types of unstructured sources; most impressively so for well-formed, closed-domain documents such as news articles or scientific publications written in English [22,25]. It has been less successful so far in processing social media content such as microblogs, known for its compact, idiosyncratic style [12]. Human computation and crowdsourcing offer an effective way to tackle these limitations [33], alongside increasingly sophisticated algorithms capitalising on the availability of huge data samples and open knowledge bases such as DBpedia and Freebase [29].

However, hybrid approaches to NER (named entity recognition) [12] that seamlessly bring together human and computational intelligence are far from being the norm. While the technology to define and deploy them is on its way - for instance, tools such as GATE already offer built-in human computation capa-

*Corresponding author. E-mail: oof1v13@soton.ac.uk.

bilities [30,6] - little is known about the overall performance of crowd-machine NER workflows and the factors that affect them. Besides various experiments reporting on task design, spam detection, and quality assurance aspects (e.g., [13,33,40]), at the moment we can only guess what features of a micropost, crowd contributor, or microtask platform will have an impact on the success of crowdsourced NER. The situation is comparable to the early stages of information extraction; once the strengths and weaknesses of particular methods and techniques had been extensively studied and understood, the research could then focus on overcoming real issues, propose principled approaches, and significantly advance the state of the art.

This paper offers an in-depth study of the factors which influence the performance of the crowd in hybrid NER approaches for microposts. We identify a set of content and crowdsourcing-related features (number of entities in a post, types of entities, skipped true-positive posts, average time spent to complete the tasks, and interaction with the user interface) and analyse their impact on the accuracy of the results and the timeliness of their delivery. We run experiments on three datasets from related literature and a fourth newly annotated corpus using CrowdFlower and our own game-with-a-purpose (GWAP) [35] called Wordsmith.[1]

An analysis of the results reveals that shorter tweets with fewer entities tend to be more amenable to microtask crowdsourcing. This applies in particular to those cases in which the text refers to single people or places, even more so when those entities have been subject to recent news or public debate on social media. Though recommended by some crowdsourcing researchers and platforms, the use of the miscellaneous as a NER category seems to confuse the contributors. However, it is well suited to identify a whole range of entities that were not explicitly targeted by the requester, from people who are less famous to partial, overlapping and what we call "*implicitly named entities*".

Structure of the paper In Section 2 we first discuss the related literature in context of the annotation of micropost data, and review existing proposals to add human and crowd computing features to the task. In Section 3 we introduce the research questions and describe the methods, experimental set-up, and data used to address them. We then present our results based on the experiment conducted and summarize the core findings in Section 7. We expect these findings to lead to the design of more advanced NER pipelines, informing the way in which tweets are chosen to be outsourced or processed by automatic tools. We make a first step in this direction by revisiting the most important lessons learnt during the experiments, framing them in the context of related literature, and discussing their implications in Section 8. We conclude with Section 9 with an overview of our contributions and an outline for future work.

Previous publications of this work This is the extended version of an eponymous paper, which was accepted for publication at ESWC2015. Compared to the original conference submission, the current paper covers a much more detailed description of the experiments, reports on additional experiments examining the same research questions as the ESWC2015 version, and expands the first study with new experiments. The new experiments look at the effect of additional detailed annotation guidelines on entity recognition accuracy and the role of sentiment analysis in crowdsourced NER. It also presents a review of a heatmap analysis which seeks to understand crowd workers behaviours in annotating entities.

Research data The results of our experiments are published as JSON-LD for further use by the research community. The download is available at `https://webobservatory.soton.ac.uk/wo/dataset/#54bd90e6c3d6d73408eb0b88`.

2. Preliminaries and related work

Several approaches have been applied to build tools for entity extraction, using rules, machine learning, or both [20]. An analysis of the state of the art in named entity recognition and linking on microposts is available in [12]. The authors also discuss a number of factors that affect precision and recall in current technology - current limitations tend to be attributed to the manner of text e.g., vocabulary words, typographic errors, abbreviations and inconsistent capitalisation, see also [14,28].

Crowdsourcing has been previously used to annotate named entities in micropost data [16]. In this study, Finin et al. used CrowdFlower and Amazon's Mechanical Turk as platforms. Crowd workers were asked to identify person (PER), location (LOC) and organisation (ORG) entities. Each task unit consisted of 5 tweets with one gold standard question, with 95% of the tweets annotated at least twice. The corpus con-

[1] `http://seyi.feyisetan.com/wordsmith`

sisted of $4,400$ tweets and 400 gold questions. Gold questions (gold data, gold standard) are questions with answers known to the task requester. This is used to evaluate worker performance and weed out spammers. A review of the results of [16] was carried out and reported in [17]. They observed annotations that showed lack of understanding of context e.g., *china* tagged as LOC when it referred to *porcelain*. They also highlighted the issue of entity drift wherein entities are prevalent in a dataset due to temporal popularity in social media. This adds to the difficulty of named entity recognition [12].

A similar approach has been used to carry out NER tasks on other types of data. Lawson et al [19] annotated $20,000$ emails using Mechanical Turk. Their approach incorporated a bonus system which allowed the payment of a bonus in addition to the base amount contingent on worker performance. The workers were also required to annotate person (PER), location (LOC), and organisation (ORG) entities. By incorporating a bonus system based on entities found and inter-annotator agreement, they were able to improve their result quality considerably. The results were used to build statistical models for automatic NER algorithms. An application in the medical domain is discussed in [39]. The crowd workers were required to identify and annotate medical conditions, medications, and laboratory tests in a corpus of $35,385$ files. They used a custom interface (just as we do with Wordsmith) and incorporated a bonus system for entities found. [37] presented a hybrid approach where expert annotators identified the presence of entities while crowd workers assigned entity types to the labels. [11] proposed a hybrid crowd-machine workflow to identify entities from text and connect them to the Linked Open Data cloud, including a probabilistic component that decides which text to be sent to the crowd for further examination. Other examples of similar systems are [7] and [30]. [30] also discussed some guidelines for crowdsourced corpus annotation (including number of workers per task, reward system, task quality approach, etc.,), elicited from a comparative study. A similar set of recommendations based on task character, human participation and motivation, and annotation quality was presented by [38].

Compared to the works cited earlier, we perform a quantitative analysis based on controlled experiments designed specifically for the purpose of exploring performance as a function of content and crowdsourcing features. The primary aim of our research is not to implement a new NER framework, but rather to understand how to design better hybrid data processing workflows, with NER as a prominent scenario in which crowdsourcing and human computation could achieve significant impact. In this context the Wordsmith game is seen as a means to outsource different types of data-centric tasks to a crowd and study their behavior, including purpose-built features for quality assurance, spam detection, and personalized interfaces and incentives.

3. Research questions

Our basic assumption was that *particular types of microposts will be more amenable to crowdsourcing than others*. Based on this premise, we identified two related research hypotheses, for which we investigated three research questions:

[H1] Specific features of microposts affect the accuracy and speed of crowdsourced entity annotation.

RQ1.1. How do the following features impact the ability of non-expert crowd contributors to recognize entities in microposts: (a) the number of entities in the micropost; (b) the type of entities in the microposts; (c) the length of micropost text; (d) the micropost sentiment?

[H2.] We can understand crowd worker preferences for NER tasks.

RQ2.1. Can we understand crowd workers preferences based on (a) the number of skipped tweets (which contained entities that could have been annotated); (b) the precision of answers; (c) the amount of time spent to complete the task; (d) the worker interface interaction (via a heatmap)?

4. Experiment design

To address these research questions we ran a series of experiments using CrowdFlower and our custom-built Wordsmith platform. We used CrowdFlower to seek help from, select, and remunerate microtask workers; each CrowdFlower job included a link to our GWAP, which is where the NER tasks were carried out. Wordsmith was used to gather insight into the features that affect a worker's speed and accuracy in annotating microposts with named entities of four types: people, locations, organisations, and miscellaneous. We describe the game in more detail in Section 6

4.1. Research data

We took three datasets from related literature, which were also reviewed by [12]. They evaluated NER tools on these corpora, while we are evaluating crowd performance. The choice of datasets ensures that our findings apply to hybrid NER workflow, in which human and machine intelligence would be seamlessly integrated and only a subset of microposts would be subject to crowdsourcing. The key challenge in these scenarios is to optimize the overall performance by having an informed way to trade-off costs, delays in delivery, and non-deterministic (read, difficult to predict) human behavior for an increase in accuracy. By using the same evaluation benchmarks we make sure we establish a baseline for comparison that allows us not only to learn more about the factors affecting crowd performance, but also about the best ways to combine human and machine capabilities.The three datasets are:
(1) **The Ritter Corpus** by [28] which consists of $2,400$ tweets. The tweets were randomly sampled, however the sampling method and original dataset size are unknown. It is estimated that the tweets were harvested around September 2010 (given the publication date and information from [12]). The dataset includes, but does not annotate Twitter *@usernames* which they argued were unambiguous and trivial to identify. The dataset consists of ten entity types.
(2) **The Finin Corpus** by [16] consists of 441 tweets which was the gold standard for a crowdsourcing annotation exercise. The dataset includes and annotates Twitter *@usernames*. The dataset annotates only 3 entity types: person, organisation and location. Miscellaneous entity types are not annotated. It is not stated how the corpus was created, however our investigation puts the corpus between August to September 2008.
(3) **The MSM 2013 Corpus**, the Making Sense of Microposts 2013 Concept Extraction Challenge dataset by [4], which includes training, test, and gold data; for our experiments we used the gold subset comprising 1450 tweets. The dataset does not include (and hence, does not annotate) Twitter *@usernames* and *#hashtags*.
(4) **The Wordsmith Corpus**, we also created and ran an experiment using our own dataset. In previous work of ours we reported on an approach for automatic extraction of named entities with Linked Data URIs on a set of 1.4 billion tweets [14]. From the entire corpus of six billion tweets, we sampled out $3,380$ English ones using *reservoir sampling*. This refers to a family of randomized algorithms for selecting samples of k items (e.g., 20 tweets per day) from a list S (or in our case, 169 days or 6 months from January 2014 to June 2014) of n items (for our dataset, over $30million$ tweets per day), where n is either a very large or an unknown number. In creating this fourth gold standard corpus, we used the NERD ontology [29] to create our annotations, e.g., a school and musical band are both sub-class-of **nerd:Organisation**, but a restaurant and museum, are sub-class-of **nerd:Location**.

The four datasets contain social media content from different time periods (2008, 2010, 2013, 2014) and have been created using varied selection and sampling methods, making the results highly susceptible to entity drift [17]. Furthermore, all four used different entity classification schemes, which we normalized using the mappings from [12]. Table 1 characterizes the data sets along the features we hypothesize might influence crowdsourcing effectivity.

4.2. Experimental conditions

We performed two experiments for each dataset; this means we evaluated $7,665$ tweets.

Condition 1
For each tweet we asked the crowd to identify four types of entities (people, locations, organisations, and miscellaneous). We elicited answers from a total of 767 CrowdFlower workers, with three assignments to each task. Each CrowdFlower job referred the workers to a Wordsmith-based task consisting of multiple tweets to be annotated. Each job was awarded 0.05 USD to annotate at least 10 tweets with no bonus incentive. We will discuss these choices in the Section 6. The workers were provided with annotation instructions detailing the various entity types and how to identify them. More details on the annotation guidelines are discussed in 6.2.

Condition 2
The second experiment condition built on the first with the same basic setup. For each tweet we asked the crowd to identify four types of entities (people, locations, organisations, and miscellaneous). Each CrowdFlower job referred the workers to a Wordsmith-based task consisting of multiple tweets to be annotated. Each job was awarded 0.05 USD to annotate at least 10 tweets with no bonus incentive. However, in the second condition, workers were presented with (i) more annotation instruction; (ii) entity type disambiguation instruction and (iii) an updated interface which pre-

Dataset overview				
Metric	Finin	Ritter	MSM2013	Wordsmith
Corpus size	441	2,400	1,450	3,380
Avg. Tweet length	98.84	102.05	88.82	97.56
Avg. @usernames	0.1746	0.5564	0.00	0.5467
Avg. #hashtags	0.0226	0.1942	0.00	0.2870
Avg. num of entities	1.54	1.62	1.47	1.72
No. PER entities	169	449	1,126	2,001
No. ORG entities	162	220	236	390
No. LOC entities	165	373	100	296
No. MISC entities	0	441	95	405
#hashtags annotated	NO	NO	NO	YES
@usernames annotated	YES	NO	NO	YES

Table 1
The four datasets used in our experiments

sented the additional instructions before annotation and inline during annotation. Effectively, we sought to understand the impact more detailed instructions would have on worker accuracy (annotation speed, precision and recall).

We also carried out basic sentiment analysis on the tweet corpora, following in the steps of [31,18]. We hypothesized that particularly polarised tweets might have an effect on the entity annotation [24]. For example, do workers annotate tweets with positive sentiments faster and more accurately compared to tweets about wars, outbreaks and tragedy. We used AlchemyAPI,[2] an external Web service providing natural language processing functionality, in order to calculate the sentiment of each tweet to be annotated.

4.3. Results and methods of analysis

The outcome of the experiments were a set of tweets annotated with entities according to the four categories mentioned earlier. We measured the execution time and compared the accuracy of the crowd inputs against the four benchmarks. By using a number of descriptive statistics to analyse the accuracy of the users performing the task, we were able to compare the precision, recall, F1 scores for entities found within and between the four datasets, as well as aggregate the performance of users in order to identify a number of distinguishing behavioural characteristics related NER tasks. Our outcomes are discussed in light of existing studies in respects to the performance of the crowd and hybrid NER workflows. For each annotation, we measured data points based on mouse movements every 10 microseconds. Each point had an x and y coordinate value which was normalized based on the worker's screen resolution. These data points were used to generate the heatmaps for our user interface analysis. For each annotation, we also recorded the time between when the worker views the tweet to when the entity details are submitted.

[2] http:www.alchemyapi.com

5. Entity types

We understood that the experiment settings would benefit from an harmonisation in the definitions of the entities. This is necessitated by the disparate nature of the entity type schemes used in the annotations of the different corpora.

5.1. Definitions and mappings

We used the NERD ontology [29] to normalise these definition even though the results were slightly different from the entity mappings adopted by [12]. Our mappings assigned *musicartist* as person (PER), distinguishing it from *musicband* which we assigned as organisation (ORG). The gains in using the nerd ontology in spite of this slight mismatch meant we could have a reference baseline when dealing with more ambiguous cases e.g., organisation-location mismatches.

5.2. Difficult cases

Organisation vs location In our preliminary experiments and gold standard creation, we noticed a number of cases that caused inter-annotator debate and disagreement. For example, given the tweets, *I am on my way to walmart* and *My local walmart made a lot of money last thanksgiving*, deciding the entity type of

Entity Mappings				
Baseline	Finin	Ritter	MSM2013	Wordsmith
Person	person -	person musicartist	per -	person -
Organisation	org -	company sportsteam	org -	organisation musicalband
Location	place -	facility geo-loc	loc -	location -
Misc	-	movie product tvshow other	-	misc

Table 2
Entity mappings across the datasets

Walmart in context becomes difficult, even for expert annotators. This extends to other classes such as museums, restaurants, universities and shopping malls.

Organisation	Location
University	Museum
Education Institution	Restaurant
-	Shopping Mall
-	Hospital

Table 3
Adopted Organisation-Location Disambiguation

Software vs organisation we also noticed a number of tweets which mentioned software which were eponymous with their parent company. For example, "*Facebook bought the photo-sharing app, Instagram*" and "*I just posted a photo on facebook :)*". The nerd ontology assigns pieces of software as a sub-class-of **nerd:Product** which maps to our miscellaneous (MISC) class. However, in cases such as these (Facebook, Instagram, Google and Twitter), we assign such entities as type organisation (ORG). For non-eponymous software or web applications e.g., *microsoft word, gmail*, these were mapped to the miscellaneous (MISC) class.

Typos, abbreviations and colloquialisms consider the tweet "*Road trip to see one of the JoBros' house w/ friends WHAT! WHAT!*". The musical band Jonas Brothers has been replaced with a collapsed *urban* form. Other examples which underscore the difficulty of the task are tweets such as "*Marry jane is the baby tho*" where "Mary" was misspelled as "Marry" (which is another name for the psychoactive drug, marijuana). Similarly, "*Jack for Wednesday*", considering the capitalisation might refer to a footballer named Jack for the football club Sheffield Wednesday, or having Jack Daniel's whiskey for Wednesday night drinks.

Nested entities consists of entities which which overlap and could potentially be annotated in multiple ways. For example, consider the following tweet from the Ritter corpus "*Gotta dress up for london fashion week and party in style !*". The correct entity in this case would be the event *london fashion week*, whereas, it workers might just annotate *London* as a location. This is also similar to identifying partial entity matches. For example, consider this tweet from the Wordsmith dataset "*Nice pass over New York City.*". The correct entity identifies New York City as opposed to a partial entity match targetting just New York.

6. Crowdsourcing approach

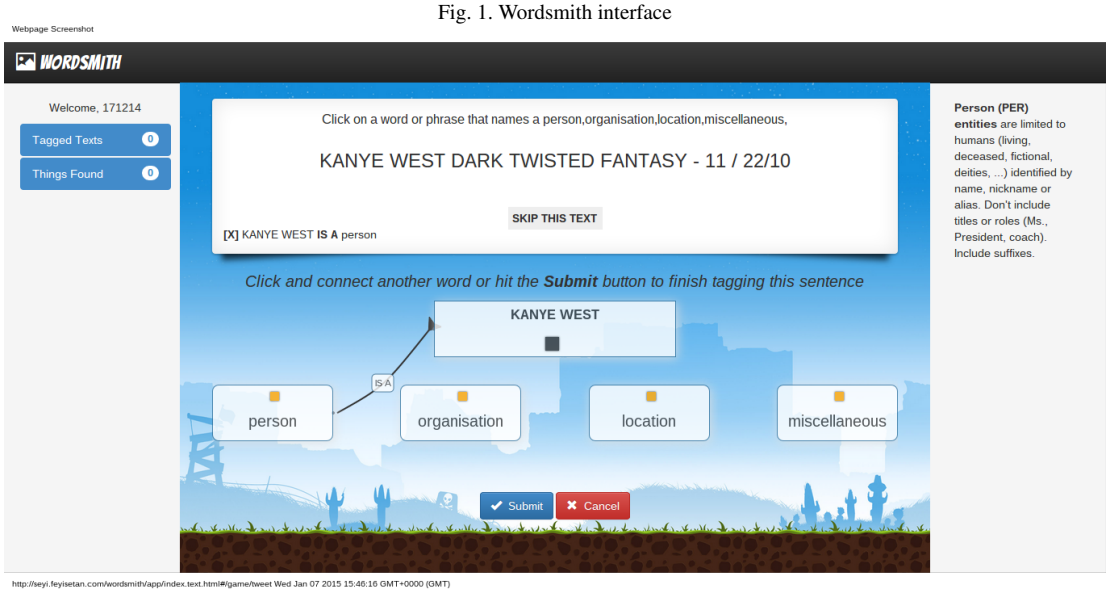
In this section, we would present an overview on our crowdsourcing approach. This includes details on our bespoke platform, our recruitment methodology using CrowdFlower, our reasons for not adopting a bonus system, our data and task model as well as our quality assurance strategy. We also elaborate on the annotation guidelines as it relates to the 2 experiment conditions, how we created our gold standard, and our approach to computing inter-annotator agreement scores.

6.1. Overview

Crowdsourcing platform: Wordsmith As noted earlier, we developed a bespoke human computation platform called *Wordsmith* to crowdsource NER tasks. The platform is designed as a GWAP and sources workers from CrowdFlower. A custom design approach was chosen in order to cater for an advanced entity recognition experience, which could not be obtained using CrowdFlower's default templates and markup language (CML). In addition, Wordsmith allowed us to set up and carry out the different experiments introduced in Section 3.

The main interface of Wordsmith is shown in Figure 1. It consists of three sections. The annotation area is at the center of the screen with sidebars for additional information. The tweet under consideration is presented at the top of the screen with each text token presented as a highlight-able span. The instruction to *'click on a word or phrase'* is positioned above the tweet, with the option to skip the current tweet below it. Custom interfaces in literature included radio buttons by [16] and span selections by [7,19,36]. We opted for a click-and-drag approach in order to fit all the annotation components on the screen (as opposed to [16]) and to cut

Fig. 1. Wordsmith interface



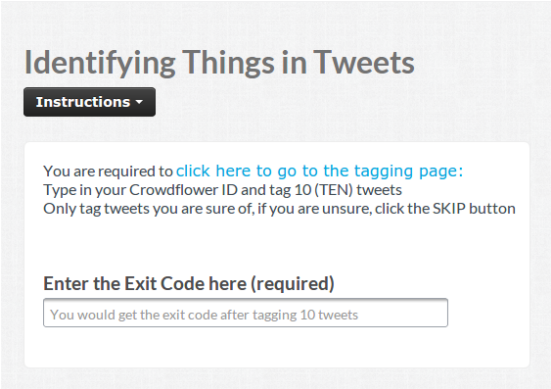
down the extra type verification step by [7]. By clicking on a tweet token(s) the user is presented with a list of connector elements representing the entity text and the entity types. Contextual information is provided in line to guide the user in making the connection to the appropriate entity type. When the type is selected, the type definition is displayed on the right hand side. The left sidebar gives an overview of the number of tweets the user has processed, and the total number of entities found. Once the worker has annotated 10 tweets, an *exit code* appears within the left side bar. This is a mechanism used to signal task completion in CrowdFlower, as we will explain in more detail later.

Recruitment We sourced the workers for our bespoke system from CrowdFlower. Each worker was invited to engage with a task as shown in Figure 2, which redirected him/her to Wordsmith. After annotating 10 tweets via the game, the worker was presented with an exit code, which was used to complete the CrowdFlower job. We recruited *Level 2 contributors*, which are top contributors who account for 36% of all monthly judgements on the CrowdFlower platform [15]. Since we were not using expert annotators, we set the judgement count at 3 answers per unit i.e., each tweet was annotated by three workers. Each worker could take on a single task unit; once starting annotating in WordSmith, they were expected to look at 10 tweets to declare the task as completed. However, they were also allowed to skip tweets (i.e., leave them unannotated) or continue engaging with the game after they reached the minimum level of 10 tweets. Independently of the actual number of posts tagged with entities, once the worker had viewed 10 of them and received the exit code, he/she receives the reward of 0.05 $.

Bonus system Unlike [19,39], we did not use any bonuses. The annotations carried out in [19] were on emails with an average length of 405.39 characters while the tweets across all our datasets had an average length of 98.24 characters. Workers in their case had the tendency to under-tag entities, a behavior which necessitated the introduction of bonus compensations which were limited and based on a worker-agreed threshold. The tasks in [39] use biomedical text, which according to them, '[is] full of jargon, and finding the three entity types in such text can be difficult for non-expert annotators'. Thus, improving recall in these annotation tasks, as opposed to shortened and more familiar text, would warrant a bonus system.

Input data and task model Each task unit refers to N tweets. Each tweet contains $x = \{0, ..., n\}$ entities. The worker's objective is to decide if the current tweet contains an entity and correctly annotate the tweet with their associated entity types. The entity types were person (PER), location (LOC), organisation (ORG), and miscellaneous (MISC). We chose our entity types based on the types mentioned in the literature of the associated datasets we used. Our task instructions encouraged workers to skip annotations they were not sure of. As we used Wordsmith as task inter-

Fig. 2. CrowdFlower interface



face, it was also possible for people to continue playing the game and contribute more, though this did not influence the payment. We report on models with adaptive rewards elsewhere [15]; note that the focus here is not on incentives engineering, but on learning about content and crowd characteristics that impact performance. To assign the total set of $7,665$ tweets to tasks, we put them into random bins of 10 tweets, and each bin was completed by three workers.

Output data and quality assurance Workers were allowed to skip tweets and each tweet was covered by one CrowdFlower job viewed by three workers. Hence, the resulting entity-annotated micropost corpus consisted of all $7,665$ tweets, each with at most three annotations referring to people, places, organisations, and miscellaneous. Each worker had two gold questions presented to them to assess their understanding of the task and their proficiency with the annotation interface. Each gold question tweet consisted of two of the entity types that were to be annotated. The first tweet was presented at the beginning, e.g., *'do you know that Barack Obama is the president of USA'* while the second tweet was presented after the worker had annotated five tweets, e.g., *'my iPhone was made by Apple'*. The workers are allowed to proceed only if they correctly annotate these two tweets.

6.2. Annotation guidelines

In each task unit, workers were required to decide whether a tweet contained entities and annotate them accordingly. As a baseline for both experiment conditions, we adopted the annotation guidelines from [16] for person (PER), organisation (ORG) and location (LOC) entity types. We also included a fourth miscellaneous (MISC) type, based on the guidelines from [28].

Experiment condition 1
Instructions were presented at the start of the CrowdFlower job, at the start via the Wordsmith interface and inline during annotation. Whenever a worker is annotating a word (or phrase), the definition of the currently selected entity type is displayed in a side bar. These instructions included the following: the task title, stated as *Identifying Things in Tweets*; an overview on the definition of entities (with a few examples); a definition of the various entity types (PER, ORG, LOC, MISC), including examples of what constitutes and does not constitute inclusion into the type categories.

Experiment condition 2
In condition 2, we provided more instructions. This included the title, stated as *Identifying Named Things in Tweets* and details on ways to handle 7 special cases. The special cases were (i) disambiguating locations such as restaurants and museums; (ii) disambiguating organisations such as universities and sport teams; (iii) disambiguating musical bands; (iv) identifying eponymous software companies; (v) dealing with nested entities by identifying the longest entities; (vi) discarding implicit unnamed entities such as hair salon, the house, bus stop; (vii) identifying and annotating *#hashtags* and *@mentions*. These instructions were placed as in *Condition 1*, with the addition of an interface update, which allowed the workers to review the additional instructions during annotation.

6.3. Gold standard creation

The gold standard used for our Wordsmith dataset was curated by 3 expert annotators among the paper authors. We manually tagged the tweet entity types using the Wordsmith platform. The Wordsmith corpus consisted of $3,380$ tweets, sampled between January 2014 to June 2014. Each tweet was annotated with the 4 designated entity types (PER, ORG, LOC, MISC). Unlike the other 3 datasets, we chose to annotate *#hashtags*. This decision was partially motivated by the nature of the dataset which had a significant number of event based *#hashtags* corresponding to the FIFA World Cup. Similarly, unlike the Ritter and MSM2013 datasets, we also annotated the *@usernames*. Our annotation choices comprised of a separation of entity types such as musical artists and musical bands as person (PER) and organisations (ORG) respectively.

6.4. Inter-annotator agreement

The inter-annotator agreement describes the degree of consensus and homogeneity in judgments among annotators [26] and is seen as a way to judge the reliability of annotated data [27]. Setting an inter-annotator threshold can enhance the precision of results from the crowd. It can be further used to shed light on our research question about crowd worker preferences for NER tasks (H2 RQ 2.1). Various scores such as the Kappa introduced by Cohen [9] have been used to calculate inter-rater agreement.

We use the approach by [5] to determine the pairwise agreement on an annotated entity text and types. Given **I** as the number of tweets in a corpus, **K** is the total number of annotations for a tweet, **H** is the number of crowd workers that annotated the tweet and **S** is the set of all entity pairs with cardinality $|S| = \binom{K}{2}$, where $k_1 = k_2 \ \forall \ \{k_1, k_2\} \in S$.

Given a tweet ***i*** and an annotated entity ***k*** where $\{k, k\} \in S$, the average agreement, A_{ik}, on the keyword ***k*** for the tweet ***i*** is given by

$$A_{ik} = \frac{n_{ik}}{\binom{H}{2}} \tag{1}$$

where n_{ik} is the number of human agent pairs that agree that annotation ***k*** is in the tweet ***i***.

Therefore, for a given tweet ***i*** the average agreement over all assigned annotations is

$$A_i = \frac{1}{|S|\binom{H}{2}} \sum_{k \in S}^{S} n_{ik} \tag{2}$$

We presented the on the average inter-annotator agreement for each corpus in the experiment in Table 12. We also presented the positive (and sometimes negative!) change in precision and recall values based on the inter-annotator thresholds in Table 13.

7. Results

7.1. Overview

The results of our experiment with condition 1 and 2 are summarised in Table 4. The first set of results in Table 4 contains precision, recall and F1 values for the four entity types for all four datasets. The results in the 2 experiment conditions (C1 and C2) show the same result patterns with matching entity types yielding the top precision and recall values. The results also show an average decrease in precision, recall and F1 scores from C1 to C2. This is in spite of the additional annotation guidelines presented in C2. This result is in line with *Myth 3* presented by [2] which states that detailed guidelines do not always yield better annotation quality. The results show highest precision scores in identifying PER entities. The only exception to this was in the Ritter dataset where the highest precision scores were in identifying LOC entities. The highest recall scores were split in between PER entities in the Ritter and MSM datasets and LOC entities in the Finin and Wordsmith datasets. However, the margins were less than 2% with a higher score recorded for PER entities in the C2 for the Finin dataset.

We also include a confusion matrix in Table 5 highlighting the entity mismatching types e.g., assigning *Cleveland* as location when it refers to the basketball team. The results show that the entity type ORG was mostly wrongly annotated as PER (in the Wordsmith dataset) and as MISC (in the Ritter dataset). The entity type LOC was most confused as the entity type ORG across all datasets (with the exception of the Ritter corpus). This occurred in both experiment condition even when more detailed instructions were given. In all dataset results, the MISC type was wrongly assigned the ORG entity type. The confusion matrix on the PER entity type was spread across all the other entity types. The Finin and Ritter showed the least confusion variance on the entity types across the two experiment conditions.

Table 6 summarises the sentiment distribution of positive, negative and neutral tweets in the different datasets. The results present the Finin, Ritter and MSM corpora as having slightly more positive than negative tweets. The Wordsmith corpus had more tweets with negative sentiments than positive. It is worth noting here that the tweets marked negative did not necessarily have to be an aggressive or abusive tweet. An example of a tweet with a negative sentiment from the Ritter dataset is "*It's the view from where I'm living for two weeks. Empire State Building = ESB. Pretty bad storm here last evening*". The next set of results in Table 7 highlights the relationship between skipped tweets and their content sentiment. The result reveals marginally that tweets with a positive sentiment were more likely to be skipped. This is inconclusive as it does not show a highly polarised set as a result of the sentiment distributions.

Table 8 like Table 7 gives further insight into the dynamics of skipped tweets. The table presents, for C1

Finin dataset						
	Condition 1: Worker annotations			Condition 2: Worker annotations		
Entity type	Precision	Recall	F1 score	Precision	Recall	F1 score
Person	**68.42**	58.96	**63.34**	43.65	**49.36**	46.33
Organisation	50.94	27.84	36.00	38.43	33.06	35.54
Location	66.14	**60.71**	63.31	**60.78**	47.67	**53.43**
Miscellaneous	-	-	-	-	-	-
Ritter dataset						
	Condition 1: Worker annotations			Condition 2: Worker annotations		
Entity type	Precision	Recall	F1 score	Precision	Recall	F1 score
Person	42.93	**69.19**	52.98	32.68	**65.72**	43.65
Organisation	28.75	39.57	33.30	27.82	42.26	33.55
Location	**67.06**	50.07	**57.33**	**62.22**	51.42	**56.31**
Miscellaneous	20.04	20.23	20.13	16.06	22.98	18.91
MSM2013 dataset						
	Condition 1: Worker annotations			Condition 2: Worker annotations		
Entity type	Precision	Recall	F1 score	Precision	Recall	F1 score
Person	**87.21**	**86.61**	**86.91**	**78.26**	**80.69**	**79.46**
Organisation	43.27	38.77	40.90	53.10	38.37	44.55
Location	60.57	67.29	63.75	49.35	59.47	53.94
Miscellaneous	10.44	29.11	15.37	5.98	30.11	9.98
Wordsmith dataset						
	Condition 1: Worker annotations			Condition 2: Worker annotations		
Entity type	Precision	Recall	F1 score	Precision	Recall	F1 score
Person	**79.23**	71.41	**75.12**	**75.95**	57.90	**65.71**
Organisation	61.07	53.46	57.01	35.97	32.30	34.04
Location	72.01	**72.91**	71.26	63.34	**65.17**	64.24
Miscellaneous	27.07	47.43	34.47	8.03	19.37	11.35

Table 4

Experiment results - Precision and Recall on the four datasets.

and C2, and across all datasets, the average number of entities present in a skipped tweet, as well as in an unskipped annotated tweet. The table also summarises, for both experiment conditions, and all datasets, the average number of characters in a skipped tweet and unskipped tweet. The tweets under consideration in the table are skipped true positive tweets i.e., tweets that were not annotated despite the presence of at least one entity. The results highlight across all datasets, that workers skipped tweets that contained more entities than the ones they annotated on average. The results present evidence that workers on average skipped longer tweets. The results were consistent across the four datasets and between the two experiment conditions. The tweet length was least significant in the MSM2013 experiment (with the number of characters between the skipped and unskipped tweet differing by less than 1 character), once again due to the comparatively well-formed nature of the dataset and the least standard deviation in the tweet lengths. The tweet length feature feature was most significant in the Ritter dataset, with workers systematically skipping tweets that were significantly longer than the average tweet length; it is worth mentioning that this corpus comprised the highest average number of characters per micropost.

More results on the skipped true-positive tweets are presented in Table 9 and Figure 3. It contains the distribution of the entities present in the posts that were left unannotated in each dataset according to the gold standard. On average across all four datasets, people tend to avoid recognizing organisations, but were more

Confusion Matrix							
Experiment Condition 1				Experiment Condition 2			
Finin dataset							
Confusion matrix (vs gold)				Confusion matrix (vs gold)			
PER	ORG	LOC	MISC	PER	ORG	LOC	MISC
78	1	7	-	498	25	67	-
1	27	5	-	52	334	27	-
1	4	84	-	2	56	431	-
-	-	-	-	-	-	-	-
Ritter dataset							
Confusion matrix (vs gold)				Confusion matrix (vs gold)			
PER	ORG	LOC	MISC	PER	ORG	LOC	MISC
765	7	26	20	2112	22	53	61
10	140	62	88	51	503	120	204
9	9	751	22	32	17	1265	30
15	46	29	217	30	106	37	500
MSM2013 dataset							
Confusion matrix (vs gold)				Confusion matrix (vs gold)			
PER	ORG	LOC	MISC	PER	ORG	LOC	MISC
3,828	25	8	7	4259	78	4	10
16	299	13	28	23	582	13	12
13	21	321	5	9	23	267	8
12	82	5	91	30	81	7	111
Wordsmith dataset							
Confusion matrix (vs gold)				Confusion matrix (vs gold)			
PER	ORG	LOC	MISC	PER	ORG	LOC	MISC
5,230	34	29	32	1750	11	12	26
93	811	30	46	50	200	21	36
25	58	1,078	8	20	68	439	0
50	113	12	718	218	48	13	102

Table 5

Experiment results - Confusion Matrix on the four datasets.

Sentiment Analysis				
Dataset	POS	NEG	NEU	UNK
Finin	41.04% (181/441)	38.10% (168/441)	20.63% (91/441)	00.23% (1/441)
Ritter	47.12% (1128/2394)	36.05% (863/2394)	15.96% (382/2394)	00.88% (21/2394)
MSM 2013	40.14% (582/1450)	34.48% (500/1450)	24.62% (357/1450)	00.76% (11/1450)
Wordsmith	36.69% (1240/3380)	46.45% (1570/3380)	16.01% (541/3380)	00.85% (29/3380)

Table 6

Sentiment Analysis - Distribution

keen in identifying locations. In the MSM2013 dataset, person entities were least skipped due to the features of the dataset discussed earlier (e.g., clear text definition, consistent capitalisation etc.). This result is also in line with those presented in Table 5 that ORG was the most misidentified entity type. This result was con-

Condition 1: Sentiment Analysis				
Dataset	POS	NEG	NEU	UNK
Finin	39.75% (64/161)	36.65% (59/161)	20.63% (38/161)	(0/161)
Ritter	38.28% (694/1813)	46.83% (849/1813)	14.62% (265/1813)	(5/1813)
MSM 2013	43.00% (562/1307)	28.84% (377/1307)	27.16% (355/1307)	(13/1307)
Wordsmith	41.98% (1508/3592)	41.25% (1482/3592)	16.31% (586/3592)	(16/3592)
Condition 2: Sentiment Analysis				
Dataset	POS	NEG	NEU	UNK
Finin	45.89% (407/888)	33.03% (293/888)	21.08% (187/888)	(1/888)
Ritter	49.67% (1895/3815)	31.66% (1208/3815)	18.03% (688/3815)	(24/3815)
MSM 2013	42.16% (729/1729)	31.52% (545/1729)	25.45% (440/1729)	(15/1729)
Wordsmith	43.25% (1150/2659)	37.57% (999/2659)	18.65% (496/2659)	(14/2659)

Table 7

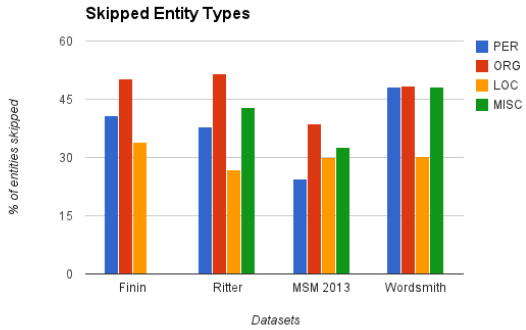
Skipped Tweets - Sentiment Analysis Distribution

Condition 1: Skipped tweets				
Dataset	**Skipped**		**Annotated**	
	Num of Entities	Tweet length	Num of entities	Tweet length
Finin	1.56	101.39	1.33	94.82
Ritter	1.42	113.05	1.35	104.22
MSM	1.49	98.74	1.30	97.11
Wordsmith	1.62	102.22	1.39	97.84
Condition 2: Skipped tweets				
Dataset	**Skipped**		**Annotated**	
	Num of Entities	Tweet length	Num of entities	Tweet length
Finin	1.51	102.44	1.20	98.99
Ritter	1.52	112.08	1.00	104.68
MSM	1.50	100.4	1.23	99.51
Wordsmith	1.61	102.70	1.39	98.14

Table 8

Experiment results - Skipped true-positive tweets

Fig. 3. Skipped tweets



sistent across both experiment conditions with crowd workers still skipping tweets with organisation entities when more instructions were given on how to disambiguate them.

Table 10 contains the average time taken for a worker to correctly identify a single occurrence of the different entity types. The results for the Finin, Ritter and MSM2013 datasets consistently present the shortest time needed corresponds to annotating locations, followed by person entities. In the Wordsmith dataset, workers correctly identified people instances in the shortest time overall, however, much longer times were taken to identify places. This result was consistent across the 2 experiment conditions with workers consistently taking shorter times to identify location and person entities. The results however note that workers took a shorter time in identifying all entity types in C2 as compared to C1. Workers took on average 1 second less to identify entities in C2. In both experiment conditions, the miscellaneous entity type took the longest time to be identified taking almost 2 seconds longer on the average as compared to location entities.

Figure 4 visualises the result of our datapoint captures via heatmaps. The results presents mouse movements concentrated horizontally along the length of the tweet text area. Much activity is also around the screen center where the entity text appears after it is clicked. The heatmaps then diverge in the lower parts of the screen which indicate which entity types were tagged. From a larger image of the interface in Figure 1, we can reconcile the mouse movements to point predominantly to PER and LOC entities in proportions which are consistent with the individual numbers presented

Condition 1: Skipped true-positive tweets				
Dataset	PER	ORG	LOC	MISC
Finin	40.91% (90/220)	50.27% (93/185)	33.83% (68/201)	-
Ritter	38.01% (631/1660)	51.57% (361/700)	26.83% (501/1867)	42.95% (847/1972)
MSM 2013	24.35% (1200/4928)	38.81% (437/1126)	30.13% (185/614)	32.58% (129/396)
Wordsmith	48.23% (4423/9170)	48.50% (796/1773)	30.35% (448/1476)	48.06% (869/1808)
Condition 2: Skipped true-positive tweets				
Dataset	PER	ORG	LOC	MISC
Finin	33.00% (435/1318)	34.83% (527/1513)	31.99% (381/1191)	-
Ritter	34.12% (1528/4478)	44.00% (898/2041)	37.11% (1305/3517)	50.67% (2067/4079)
MSM 2013	23.57% (1633/6928)	28.09% (545/1940)	30.67% (196/639)	35.99% (203/564)
Wordsmith	50.86% (2952/5804)	44.83% (473/1055)	35.22% (329/934)	50.05% (514/1027)

Table 9

Experiment results - Skipped true-positive tweets

Condition 1: Avg. Annotation Time				
Dataset	PER	ORG	LOC	MISC
Finin	9.54	12.15	8.91	-
Ritter	9.69	10.05	9.35	10.88
MSM	9.54	10.77	8.70	10.35
Wordsmith	8.06	8.50	9.56	9.48
Condition 2: Avg. Annotation Time				
Dataset	PER	ORG	LOC	MISC
Finin	7.20	7.05	6.94	-
Ritter	8.70	9.01	8.65	10.22
MSM	7.73	8.75	7.76	9.69
Wordsmith	6.88	6.79	6.97	8.72

Table 10

Experiment results - Average accurate annotation time

Average Position of First Entity		
Dataset	Gold Entity	User Entity
Finin	16.91	22.93
Ritter	34.56	22.81
MSM 2013	35.61	24.77
Wordsmith	14.68	21.33

Table 11

Experiment results - Average Position of First Entity

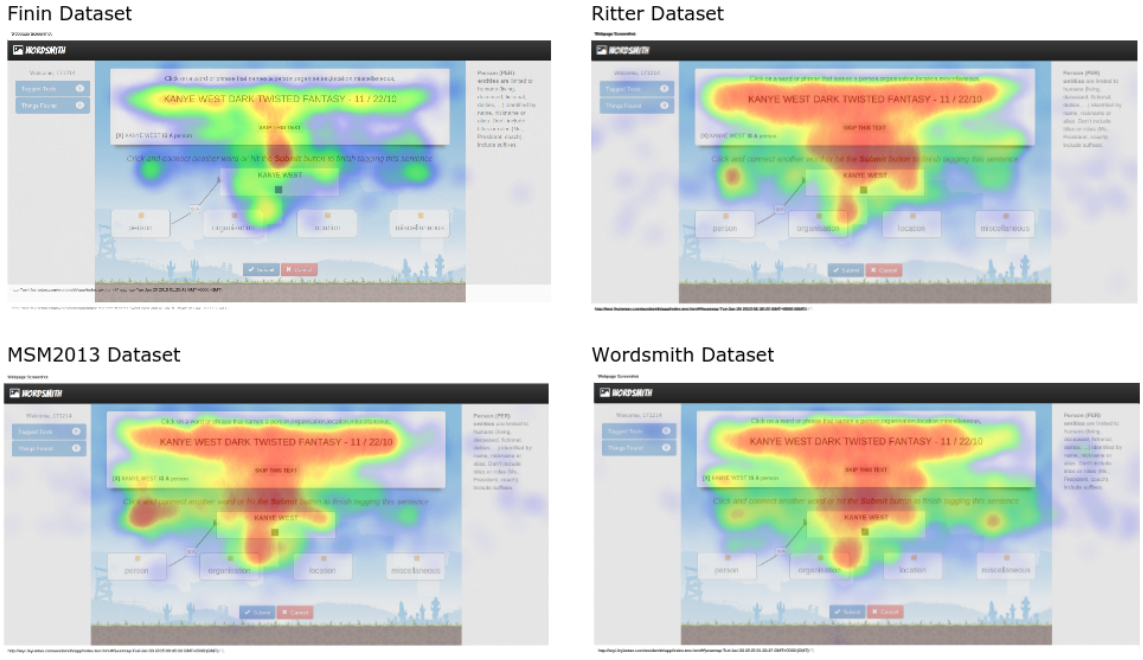
in Table 4. A corollary to the visualisation presented in the heatmaps is the result outlined in Table 11. The results contain the average position of the first entity in the dataset gold standard and the average position of the first entity annotated by the workers. From the results we note that although the average positions in the gold standards vary from the 14th character in the Wordsmith dataset to the 35th character in the MSM dataset, the average worker consistently tagged the first entity around the 21st to 24th character mark. This result was consistent across all the four dataset and in variance with the results from the gold standards. We would shed more light into this in the discussion section.

Table 12 summarises the average inter-annotator scores across the four datasets. It points out an average of 35% agreement in the Ritter, MSM and Wordsmith corpora. The Finin corpus however has an inter-annotator score of 60%. A follow up to the result is presented in Table 13. The results highlight the changes in precision and recall scores obtained by assigning a minimum inter-annotator threshold. The results detail scores obtained by setting a baseline of at least 2 workers agreeing on an annotation and at least 3 workers agreeing on an annotation.

An agreement threshold of 2 workers was beneficial for the precision of identifying all the entity types across all datasets. This effect was strongest in the Wordsmith dataset where a minimum threshold of 2 raised the precision scores of identifying organisations by 20%. The least significance of the inter-annotator threshold was in identifying miscellaneous entity types in the MSM dataset where the precision score moved up by barely 0.5%. The recall values for identifying locations were the most enhanced by setting a threshold agreement of at least 2 workers. The raise in recall also showed the least gain in the miscellaneous entity types in the MSM dataset.

Increasing the agreement threshold to at least 3 workers showed a further surge consistent with the results from setting a threshold of 2. The highest pre-

Fig. 4. Wordsmith Heatmaps across the 4 datasets



cision scores are also from the Wordsmith dataset in identifying organisations which had a boost of 30%. Precision scores in the MSM and Ritter datasets also went up over 20% by setting the inter-annotator worker threshold to a minimum of 3. As with the results presented in the previous paragraph, the lowest precision and recall score enhancements came from annotating miscellaneous entity types in the MSM dataset.

Average Inter-annotator Agreement				
Dataset	Finin	Ritter	MSM	Wordsmith
Score	60.29%	36.11%	35.15%	35.00%

Table 12

Experiment results - Average Inter-annotator Agreement

7.2. Summary of findings

7.2.1. Overview

The low performance values for the Ritter dataset can be attributed in part to the annotation schema (just as in [12]). For example, the Ritter gold corpus assigns the same entity type *musicartist* to single musicians and group bands. More significantly, the dataset does not annotate Twitter *@usernames* and *#hashtags*. Considering that most *@usernames* identify people and organisations, and the corpus contained 0.55 *@usernames* per tweet (as listed in Table 1), it is not surprising that scores are rather low. The result also reveals high precision and low confusion in annotating locations, while the greatest ambiguities come from annotating miscellaneous entities.

The Finin dataset has higher F1 scores across the board when compared to the Ritter experiments. The dataset did not consider any MISC annotations and although it includes *@usernames* and *@hashtags*, only the *@usernames* are annotated. Here again, the best scores were in the identification of people and places.

For the MSM2013 dataset highest precision and recall scores were achieved in identifying PER entities. However, it is important to note that this dataset (as highlighted in Table 1) contained, on average, the shortest tweets (88 characters). In addition, the URLs, *@usernames* and *#hastags* were anonymized as _URL_, _MENTION_ and _HASHTAG_, hence the ambiguity arising from manually annotating those types was removed. Furthermore, the corpus had a disproportionately high number of PER entities (1,126 vs. just 100 locations). It also consisted largely of clean, clearly described, properly capitalised tweets, which could have contributed to the precision. Con-

Finin dataset				
	Inter Annotator ≥ 2		Inter Annotator ≥ 3	
Entity	Precision	Recall	Precision	Recall
PER	2.77	4.69	2.12	4.61
ORG	7.65	3.33	9.17	5.37
LOC	**8.74**	**9.17**	**12.45**	**13.01**
MISC	-	-	-	-
Ritter dataset				
	Inter Annotator ≥ 2		Inter Annotator ≥ 3	
Entity	Precision	Recall	Precision	Recall
PER	5.11	5.17	9.83	7.65
ORG	**14.60**	4.62	**22.85**	5.74
LOC	11.58	**6.92**	16.46	**10.52**
MISC	14.35	3.79	22.37	2.62
MSM2013 dataset				
	Inter Annotator ≥ 2		Inter Annotator ≥ 3	
Entity	Precision	Recall	Precision	Recall
PER	5.38	4.53	6.37	6.10
ORG	**15.33**	3.66	**21.18**	4.12
LOC	11.67	**8.52**	14.72	**9.99**
MISC	0.49	1.12	0.60	-3.34
Wordsmith dataset				
	Inter Annotator ≥ 2		Inter Annotator ≥ 3	
Entity	Precision	Recall	Precision	Recall
PER	11.30	**9.09**	14.16	**13.76**
ORG	**20.49**	2.34	**29.69**	0.77
LOC	10.15	7.07	13.28	10.06
MISC	10.68	2.64	31.97	0.56

Table 13

Inter Annotator Deltas - Change in precision and recall values based on different inter-annotator thresholds

sistent with the results above, the highest scores were in identifying PER and LOC entities, while the lowest one was for those entities classified as miscellaneous.

Our own *Wordsmith dataset* achieved the highest precision and recall values in identifying people and places. Again, crowd workers had trouble classifying entities as MISC and significant noise hindered the annotation of ORG instances. A number of ORG entities were misidentified as PER and an equally high number of MISC examples were wrongly identified as ORG. The Wordsmith dataset consisted of a high number of *@usernames* (0.55 per tweet) and the highest concentration of *#hashtags* (0.28 per tweet).

Disambiguating between ORG and LOC types remained challenging across all datasets as evidenced in the confusion matrices in Table 5. Identifying locations such as *London* was a trivial task for contributors, however, entities such as museums, shopping malls, and restaurants were alternately annotated as either LOC or ORG. Disambiguating tech organisations was not trivial either - that is, distinguishing entities such as Facebook, Instagram, or Youtube as Web applications or independent companies without much context. In the Wordsmith dataset, however, PER, ORG, and MISC entity tweets were skipped with equal likelihood. This is likely due to a high number of these entities arising from *@usernames* and *#hashtags*, as opposed to well-formed names. As noted earlier, this was a characteristic of this dataset, which was not present in the other three.

7.2.2. Analysis of tweet features

We now discuss our results in light of H1 RQ1.1 which states that specific features of microposts affect the accuracy and speed of crowdsourced entity annotation. We focus on four main features (a) the number of entities in the micropost; (b) the type of entities in the microposts; (c) the length of micropost text; (d) the micropost sentiment.

Number of entities

From the results in 8 we see that the number of entities in a tweet affect the likelihood of annotation by a worker. We note that workers were more likely to annotate tweets which had fewer entities than the dataset average as contained in Table 1. This is further seen in the lower recall scores (as compared to precision) in Table 4; workers are more likely to annotate one entity in a tweet, or completely ignore tweets which have more entities than the dataset average.

Entities types

Table 9 and Figure 3 give details on skipped true positive tweets and the corresponding entity distributions. The table indicates for each dataset the total entity type encounters by the crowd workers and how many were skipped. For the first experiment condition C1 with the baseline annotation guidelines, workers skipped tweets that contained ORG entities with the highest frequency. Comparing this with our dataset overview in Table 1, we observe that the ORG type was not the most occurring entity type in any of the datasets, yet it was the most skipped. The next most skipped entity type was the MISC entity type in the MSM and Ritter corpora (there were no MISC annotations in the Finin gold standard). The Wordsmith dataset had the PER,

ORG and MISC entity types skipped with equal frequency. For the Wordsmith dataset, as discussed earlier, this can be attributed also to entities arising from *@usernames* and *#hashtags*. The other datasets either exclude them or do not annotate them in their gold standards.

In the second experiment condition C2, in which workers were given further instructions on how to disambiguate entity types such as restaurants and museums as LOC; and universities, sport teams and musical bands as ORG, workers were then less likely to skip this entity type. Even though this did not raise precision and recall scores (as seen in Table 4), workers did not skip the ORG entity types as often as they did without the instructions. 3 of the 7 extra instructions explained in some form how to identify ORG entities and this likely contributed to them being skipped less. In C2, the MISC entity type was the most skipped on the average. People-related tweets were skipped more in the Finin and Wordsmith dataset, but this is a function of the high number of entities of this type (see also Table 1) rather than an indicator of crowd behaviour. The MSM dataset had a high number of PER entities, however, these were not skipped as the tweets were well mainly formed well structured texts e.g., quotes with the author attribution at the end.

Micropost text length

Table 8 reveals that workers prefer tweets with fewer characters. The Ritter dataset with a mean tweet length of 102 characters had workers annotating posts which hovered slightly about this average length. The MSM2013 dataset had the shortest tweets with an average length of 88 characters, however, workers were willing to annotate annotate tweets with up to 9 characters above the corpus average. The Finin and Wordsmith datasets both had tweets with an average length of 98 characters with workers annotating similarly around this average point.

These results are reinforced in C2 with workers annotating tweets in the 98-99 character length set and discarding tweets over 100 characters. This is asides the Ritter dataset which had an overall set of longer tweets. From this we observe that regardless of the dataset (such as the MSM dataset with an average length of 88 characters), workers would be willing to annotate up to a certain threshold before they start skipping.

Micropost sentiment

Our experiments indicate marginally that tweets with a positive sentiment were more likely to be skipped. This is inconclusive as it does not show a polarised set as a result of the sentiment distributions. It might be possible to study the effect of tweet sentiment in annotations by carrying out granular sentiment analysis, categorising tweets as nervous, tense, excited, depressed, rather than assigning the generic positive, negative and neutral labels. Sentiment features might also be prominent in a dataset that features deleted tweets, flagged tweets or reported tweets. Other potential classes might be tweets posted to celebrities or tweets during sporting events and concerts.

7.3. Analysis of behavioral features of crowd workers

We now discuss our results in light of H2 RQ2.1, which states that we can understand crowd workers preferences based on (a) the number of skipped tweets (which contained entities that could have been annotated); (b) the precision of answers; (c) the amount of time spent to complete the task; and (d) the worker interface interaction.

Number of skipped tweets

Tables 8, 9, and 7 give insights into the skipped tweets. The results show that the number of entities and the length of the tweet were two factors that contributed to the likelihood of a skipped tweet. At this time we cannot present conclusive remarks on the effect of the tweet sentiment on a workers probability of annotating it.

Accuracy of answers

From the results in Table 4 we note that the crowd workers were better at identifying PER and LOC entities, and poor at characterizing MISC entity types. Table 5 gives further insights into the mismatching between organisation and locations (e.g., restaurants), organisations and persons (e.g., musical bands) and organisations and miscellaneous entities.

Amount of time spent to complete the task

As shown in Table 10 locations and people are quickly identified. In addition, the tagging speed goes up with an expansion in annotation guidelines (although the accuracy remains constant or even declines slightly). Tweets with MISC entities took the longest time to be annotated.

Worker interface interaction

We presented the findings from our heatmap datapoints in the result section and visualised them in Figure 4.

Table 11 further shows us that workers tend to start annotating around a specific start point. In our experiments, we discovered that regardless of the dataset, workers started l entities that occurred around the 21st to 24th character. The Finin and Wordsmith dataset however had much lower start points in their gold standard (after 15 characters) while the Ritter and MSM corpora had much higher ones (after 35 characters). We took into consideration the responsive nature of the interface which could have presented the annotation text slightly different on varying screen resolutions and with screen resizing, and ensured that the micropost texts were presented in the same way on various screens.

8. Discussion

In this final section we assimilate our results into a number of key themes and discuss their implications on the prospect of hybrid NER approaches that combine automatic tools with human and crowd computing.

Crowds can identify people and places, but more expertise is needed to classify miscellaneous entities Our analysis clearly showed that microtask workers are best at spotting locations, followed by people, and finally with a slightly larger gap, organisations. When no clear instructions are given, that is, when the entity should be classified as MISC, the accuracy suffers dramatically. Assigning entities as organisations seems to be cognitively more complex than persons and places, probably because it involves disambiguating their purpose in context e.g., universities, restaurants, museums, shopping malls. Many of these entities could also be ambiguously interpreted as products, brands, or even locations, which also raises the question of more refined models to capture diverse viewpoints in annotation gold standards [1]. To improve the crowd performance, one could imagine interfaces and instructions that are bespoke for this type of entities. However, this would assume the requester has some knowledge about the composition of his corpus and can identify problematic cases. A similar debate has been going on in the context of GWAPs, as designers are very restricted in assigning questions to difficulty levels without pre-processing them [32]. One option would be to try out a multi-step workflow in which entity types are empirically straightforward to annotate are solved by 'regular' workers, while miscellaneous and other problematic cases are only flagged and treated differently - be that by more experienced annotators, via a higher number of judgements, or otherwise.

Crowds perform best on recent data, but remember people All four analyzed datasets stem from different time periods (Ritter from 2008, Finin from 2010, MSM from 2013, and Wordsmith from 2014). Most significantly one can see that there is a consistent build-up of the F1 score the more recent the dataset is, even if the difference is only a couple of months as between the MSM and the Wordsmith cases. We interpret that the more timely the data, the better the performance of crowd workers, possibly due to the fact that newer datasets are more likely to refer to entities that gained public visibility in media and on social networks in recent times and that people remember and recognize easily. This concept known as entity drift was also highlighted by [12,17]. The only exception for this is the PER entity type, which was the most accurate result for the MSM dataset. However, in order to truly understand this phenomenon we would need more extended experiments, focusing particularly on people entities, grounded in cognitive psychology and media studies [8,23].

Partial annotations and annotation overlap The experiments showed a high share of partial annotations by the workers. For example, workers annotated *london fashion week* as *london* and *zune hd* as *zune*. Other partial annotations stemmed from identifying a person's full name, e.g., *Antoine De Saint Exupery* was tagged by all three annotators as *Antoine De Saint*. Overlapping entities occurred when a text could refer to multiple nested entities e.g., *berlin university museum* referring to the university and the museum and *LPGA HealthSouth Inaugural Golf Tournament* which was identified as an organisation and an event. These findings call for richer gold standards, but also for more advanced means to assess the quality of crowd results to reward partial answers. Such phenomena could also signal the need for more sophisticated microtask workflows, possibly highlighting partially recognized entities to acquire new knowledge in a more targeted fashion, or by asking the crowd in a separate experiment to choose among overlaps or partial solutions.

Spotting implicitly named entities thanks to human reasoning Our analysis revealed a notable number of entities that were not in the gold standard, but were

picked up by the crowd. A manual inspection of these entities in combination with some basic text mining has shown that the largest set of these entities suggest that human users tend to spot unnamed entities (e.g., *prison* or *car*), partial entities (e.g., *apollo* versus *the apollo*), overlapping entities (e.g., *london fashion week* versus *london*), and hashtags (e.g., *#WorldCup2014*). However, the most interesting case were the ones we call *implicitly named entities*. Examples such as *hair salon*, *last stop*, *in store*, or *bus stop* give evidence that the crowd is good at spotting phrases that refer to real named entities implicitly depending on the context of the post's author or a person or event this one refers to. In many cases, the implicit entities found are contextualised within the micropost message, e.g., *I'll get off at the stop after Waterloo*. This opens up interesting directions for future analysis that focus only on those implicit entities together with features describing their context in order to infer the actual named entity in a human-machine way. By combining text mining and content analysis techniques, it may be possible to derive new meaning from corpora such as those used within this study.

Closing the entity linking loop for the non-famous Crowd workers have shown good performance in annotating entities that were left out by the gold standards and presented four characteristic classes of such entities (unnamed entities, partial entities, overlapping entities, and hashtags). We observed a fifth class that human participants mark as entities, which refer to non-famous, less well-known people, locations, and organisations (e.g., the name of a person who is not a celebrity). This is an important finding for hybrid entity extraction and linking pipelines, which can benefit from the capability to generate new URIs for yet publicly unknown entities. This can play an important role in data journalism [21].

Wide search, but centred spot Our heatmap analysis indicated that we had a very wide view along the text axis, and a consistent pattern that the likelihood of annotating in the center is higher even though they seem to search over the entire width of the text field. This correlates with statistics about the average position of the first annotation, which remained constant in the user annotations as compared to the varying positions in the gold standard. Workers started off by annotating entities at the beginning of the tweet then around the middle of the tweet before the tagging recall dropped. This might mean that people are more likely to miss out on annotating entities on the right edges of the interface or at the end of the text. A resolution could be to centralize the textbox and make it less wide hence constraining the worker's field of vision as opposed to [16] where workers were required to observe vertically to target entities.

Useful guidelines are an art Our study seems to indicate that additional instructions do not always produce better tagging quality. We noted, however, that it has the following effects (i) it speeds up the annotation process as we noted that workers spent less time on the average annotating entities; and (ii) it makes people more willing to undertake choice-based work - tweets with ORG entities were less skipped after the introduction of more detailed guidelines. However, this did not affect the accuracy scores, which were in fact reduced in a few places. The new guidelines did not remove worker bias towards identifying implicit unnamed entities. Workers continued to tag concepts such as room, gym and on the road as entities even when the instructions tried to discourage them to do so. While giving effective feedback is an ongoing research problem in crowdsourcing, one approach which we could investigate more is crowd-based feedback and crowd sociality, using synchronous work by workers who are completing tasks in the same time. A previous study we carried out [15] points out that crowd workers appreciate features which offer continuous feedback mechanisms and a view into how other workers are performing with the task. Another interesting question would be if we could leverage the efforts people invested in tagging things we were not looking for. While it is clear that crowdsourcing, at least on paid microtask platforms, is goal-driven and that the requester is the one setting the goals, it might make sense to consider models of co-creation and task autonomy, in which as the tasks are being completed, the requester takes into account the feedback and answers of the crowd and adjusts the goals of the project accordingly. Literature on motivation tells us that people perform best when they can decide what they are given the freedom to choose what they contribute, how, and when, and when they feel they are bringing in their best abilities [10]. These aspects might not be at the core of CrowdFlower and others, which focus on extrinsic motivation and rewards, but they are nevertheless important and could make experiments more useful in several ways.

9. Conclusion and future work

In this paper we studied an approach to finding entities within micropost datasets using crowdsourced methods. Our experiments, conducted on four different corpora, revealed a number of crowd characteristics with respect to their performance and behaviour of identifying different types of entities. In terms of the wider impact of our study, we consider that our findings will be useful for streamlining and improving hybrid NER workflows, offering an approach that allows corpora to be divided up between machine and human-led workforces, depending on the types and number of entities to be identified or the length of the tweets. Future work in this area includes (i) devising automated approaches to determining when best to select human or machine capabilities; (ii) examining *implicitly named entities* in order to develop methods to identify and derive message-related context and meaning; as well as (iii) looking into alternative ways to engage with contributors using real-time crowdsourcing, crowd feedback, multi-steps workflows involving different kinds of expertise to improve tagging performance for organizations and other ambiguous entities, and giving the contributors more freedom and autonomy in the annotation process.

References

[1] L. Aroyo and C. Welty. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013.

[2] Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, mar 2015.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. *Dbpedia: A nucleus for a web of open data*. Springer, 2007.

[4] A. E. C. Basave, A. Varga, M. Rowe, M. Stankovic, and A. Dadzie. Making sense of microposts (# msm2013) concept extraction challenge. In *# MSM*, pages 1–15, 2013.

[5] Plaban Kr Bhowmick, Pabitra Mitra, and Anupam Basu. An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 58–65. Association for Computational Linguistics, 2008.

[6] Kalina Bontcheva, Leon Derczynski, and Ian Roberts. Crowdsourcing named entity recognition and entity linking corpora. *Handbook of Linguistic Annotation. Springer*, 2014.

[7] K. Braunschweig, M. Thiele, J. Eberius, and W. Lehner. Enhancing named entity extraction by effectively incorporating the crowd. In *BTW Workshops'13*, pages 181–195, 2013.

[8] Yu Cheng, Zhengzhang Chen, Jiang Wang, Ankit Agrawal, and Alok Choudhary. Bootstrapping active name disambiguation with crowdsourcing. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1213–1216. ACM, 2013.

[9] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37, 1960.

[10] E.L. Deci and R.M. Ryan. *Intrinsic Motivation and Self-Determination in Human Behavior*. Perspectives in Social Psychology. Springer, 1985.

[11] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st international conference on World Wide Web*, pages 469–478. ACM, 2012.

[12] L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak, and K. Bontcheva. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49, 2015.

[13] D. E. Difallah, G. Demartini, and P. Cudré-Mauroux. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In *CrowdSearch*, pages 26–30, 2012.

[14] O. Feyisetan, E. Simperl, R. Tinati, M. Luczak-Roesch, and N. Shadbolt. Quick-and-clean extraction of linked data entities from microblogs. In *Proceedings of the 10th International Conference on Semantic Systems*, SEM '14, pages 5–12. ACM, 2014.

[15] O. Feyisetan, E. Simperl, and M. Van Kleek. Improving paid microtasks through gamification and adaptive furtherance incentives. In *Proceedings of the 24th international conference on World wide web*. International World Wide Web Conferences Steering Committee, 2015.

[16] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. 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*, pages 80–88. Association for Computational Linguistics, 2010.

[17] H. Fromreide, D. Hovy, and A. SÃÿgaard. *Crowdsourcing and annotating NER for Twitter #drift*. European language resources distribution agency, 2014.

[18] Alec Go, Lei Huang, and Richa Bhayani. Twitter sentiment analysis. *Entropy*, 17, 2009.

[19] N. Lawson, K. Eustice, M. Perkowitz, and M. Yetisgen-Yildiz. 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*, pages 71–79. Association for Computational Linguistics, 2010.

[20] X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 359–367. ACL, 2011.

[21] Markus Luczak-Rösch and Ralf Heese. Linked data authoring for non-experts. In *LDOW*, 2009.

[22] M. Marrero, S. Sanchez-Cuadrado, J. M. Lara, and G. Andreadakis. Evaluation of named entity extraction systems. *Advances in Computational Linguistics, Research in Computing Science*, 41:47–58, 2009.

[23] Einat Minkov, Richard C Wang, and William W Cohen. Extracting personal names from email: applying named entity

recognition to informal text. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 443–450. Association for Computational Linguistics, 2005.

[24] Robert Morris. Crowdsourcing workshop: the emergence of affective crowdsourcing. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems. ACM*, 2011.

[25] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[26] Stefanie Nowak and Stefan Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566. ACM, 2010.

[27] Rohan Ramanath, Monojit Choudhury, Kalika Bali, and Rishiraj Saha Roy. Crowd prefers the middle path: A new iaa metric for crowdsourcing reveals turker biases in query segmentation. In *ACL (1)*, pages 1713–1722, 2013.

[28] A. Ritter, S. Clark, O. Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

[29] G. Rizzo and R. Troncy. Nerd: evaluating named entity recognition tools in the web of data. 2011.

[30] M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proc. LREC*, 2014.

[31] Hassan Saif, Yulan He, and Harith Alani. Semantic sentiment analysis of twitter. In *The Semantic Web–ISWC 2012*, pages 508–524. Springer, 2012.

[32] Elena Simperl, Roberta Cuel, and Martin Stein. Incentive-centric semantic web application engineering. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 3(1):1–117, 2013.

[33] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics, 2008.

[34] R. Usbeck, A. Ngomo, M. Röder, D. Gerber, S. A. Coelho, S. Auer, and A. Both. Agdistis-graph-based disambiguation of named entities using linked data. In *The Semantic Web–ISWC 2014*, pages 457–471. Springer, 2014.

[35] L. von Ahn and L. Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8):58–67, August 2008.

[36] R. Voyer, V. Nygaard, W. Fitzgerald, and H. Copperman. A hybrid model for annotating named entity training corpora. In *Proceedings of the fourth linguistic annotation workshop*, pages 243–246. Association for Computational Linguistics, 2010.

[37] Robert Voyer, Valerie Nygaard, Will Fitzgerald, and Hannah Copperman. A hybrid model for annotating named entity training corpora. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 243–246, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[38] A. Wang, C. D. V. Hoang, and M. Kan. Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47(1):9–31, 2013.

[39] M. Yetisgen-Yildiz, I. Solti, F. Xia, and S. R. Halgrim. Preliminary experience with amazon's mechanical turk for annotating medical named entities. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 180–183. Association for Computational Linguistics, 2010.

[40] M. Yuen, I. King, and K. Leung. A survey of crowdsourcing systems. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 766–773. IEEE, 2011.