Towards Hybrid NER: A Study of Content and Crowdsourcing-Related Performance Factors

Oluwaseyi Feyisetan(✉), Markus Luczak-Roesch, Elena Simperl,
Ramine Tinati, and Nigel Shadbolt

University of Southampton, Southampton, UK
oof1v13@soton.ac.uk
http://www.sociam.org

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, 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 them 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 [2]. 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 [20]. State-of-the-art technology in entity recognition achieves near-human performance for many types of unstructured sources, and most impressively so for well-formed, closed-domain documents such as news articles or scientific publications written in English [14,15]. It has been less successful so far in processing social media content such as microblogs, known for its compact, idiosyncratic style [6]. Human computation and crowdsourcing offer an effective way to tackle these limitations [19], alongside increasingly sophisticated algorithms capitalising on the availability of huge data samples and open knowledge bases such as DBpedia and Freebase [17].

F. Gandon et al. (Eds.): ESWC 2015, LNCS 9088, pp. 525–540, 2015.
DOI: 10.1007/978-3-319-18818-8_32

However, such hybrid approaches to NER (named entity recognition) [6] 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 capabilities [18] - 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., [7,19,24]), 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 can then focus on overcoming real issues, propose principled approaches, and significantly advance the state of the art.

This paper is a first in-depth study that examines 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) [21] 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 entities, even more so when these 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, they are 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*".

The remainder of this paper is structured as follows: 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 Sect. 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 finally discuss the core findings. We expect them 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. The results of our experiments are published as JSON-LD for further use by the research community.[2]

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

[2] Download 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 [13]. An analysis of the state of the art in named entity recognition and linking on microposts is available in [6]. 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 [8,16].

Crowdsourcing has been previously used to annotate named entities in micropost data [10]. 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 consisted of 4, 400 tweets and 400 gold questions. A review of the results of [10] was carried out and reported in [11]. 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 [6].

A similar approach has been used to carry out NER tasks on other types of data. Lawson et al. [12] annotated 20, 000 emails using Mechanical Turk. 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 [23]. 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. Reference [5] 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 [4,18]. Reference [18] 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.

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 and Experiment Design

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?

[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)?

RQ2.2. How do these four worker-related dimensions correlate with the content features from RQ1.1?

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 Sect. 4

Research data. We took three datasets from related literature, which were also reviewed by [6]. 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 [16] 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 [6]). 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 [10] 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 Making Sense of Microposts 2013 Concept Extraction Challenge dataset by [3], 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.*

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 [8]. 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 $30 million$ 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 [17] 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 [11]. Furthermore, all four used different entity classification schemes, which we normalized using the mappings from [6]. Table 1 characterizes the data sets along the features we hypothesize might influence crowdsourcing effectivity.

Experimental conditions. We performed one experiment for each dataset, which adds up to $7,665$ tweets. 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 with no bonus. We will discuss these choices in the next section.

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

Table 1. The four datasets used in our experiments

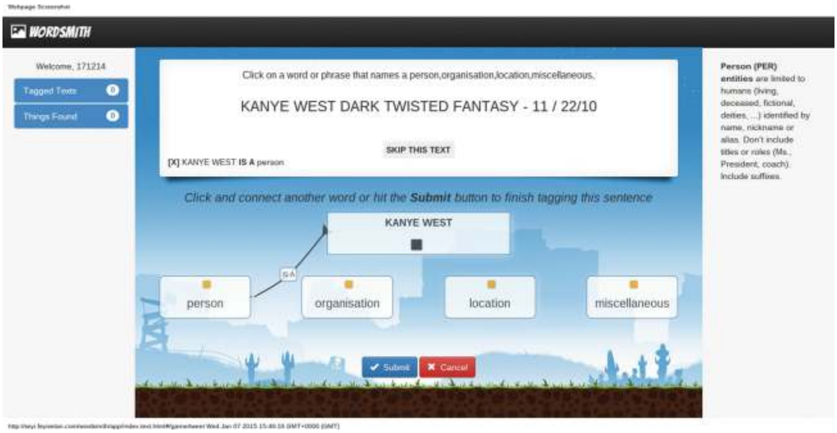
Dataset overview				
Metric	Finin	Ritter	MSM2013	Wordsmith
Corpus size	441	2,400	1,450	3,380
Avg. Tweet length	98.84	102.05	88.82	97.56
Avg. @usernames	0.1746	0.5564	0.00	0.5467
Avg. #hashtags	0.0226	0.1942	0.00	0.2870
No. PER entities	169	449	1,126	2,001
No. ORG entities	162	220	236	390
No. LOC entities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

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 ms. 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.

4 Crowdsourcing Approach

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 Sect. 3.

The main interface of Wordsmith is shown in Fig. 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 [10] and span selections by [4,12,22]. We opted for a click-and-drag approach in order to fit all the annotation components on the screen (as opposed to [10]) and to cut down the extra type verification step by [4]. By clicking on a tweet



Click on a word or phrase that names a person,organisation,location,miscellaneous.

KANYE WEST DARK TWISTED FANTASY - 11 / 22/10

[X] KANYE WEST **IS A** person

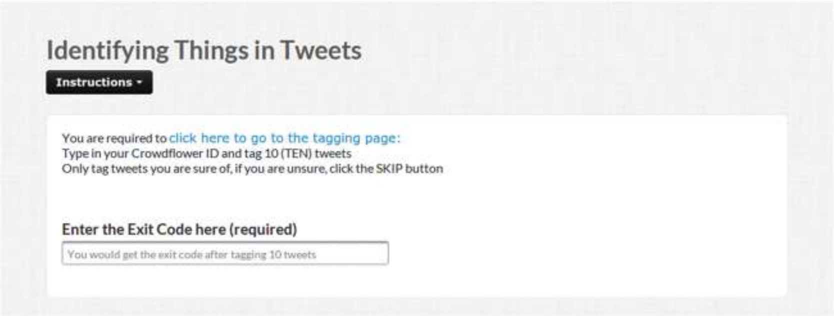
*Click and connect another word or hit the **Submit** button to finish tagging this sentence*

Person (PER) entities are limited to humans (living, deceased, fictional, deities, ...) identified by name, nickname or alias. Don't include titles or roles (Ms., President, coach). Include suffixes.

Fig. 1. Wordsmith interface

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 Fig. 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 [9]. Since we were not using expert annotators, we set the judgement count at 3 answers per unit i.e.,



Identifying Things in Tweets

You are required to click here to go to the tagging page:
Type in your Crowdflower ID and tag 10 (TEN) tweets
Only tag tweets you are sure of, if you are unsure, click the SKIP button

Enter the Exit Code here (required)

You would get the exit code after tagging 10 tweets

Fig. 2. CrowdFlower interface

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 $.

Unlike [12,23], we did not use any bonuses. The annotations carried out in [12] 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 [23] 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 interface, 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 [9]; 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.

Annotation guidelines. In each task unit, workers were required to decide whether a tweet contained entities and annotate them. We adopted the annotation guidelines from [10] for person (PER), organisation (ORG) and location (LOC) entity types. We also included a fourth miscellaneous (MISC) type, based on the guidelines from [16]. 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.

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.

5 Results

5.1 Analysis of Micropost Features

The first set of results in Table 2 shows precision, recall and F1 values for the four entity types for all four datasets. We also include a confusion matrix highlighting the entity mismatching types e.g., assigning *Cleveland* as location when it refers to the basketball team. The low performance values for the Ritter dataset can be attributed in part to the annotation schema (just as in [6]). 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 shown in Table 1), it is not surprising that scores are rather low. The result also shows high precision and low confusion in annotating location entities, while the greatest ambiguities come from annotating miscellaneous ones.

The results for the Finin dataset show 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 the results show the highest precision and recall scores in identifying PER entities. However, it is important to note that this dataset (as shown 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. Consistent 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.

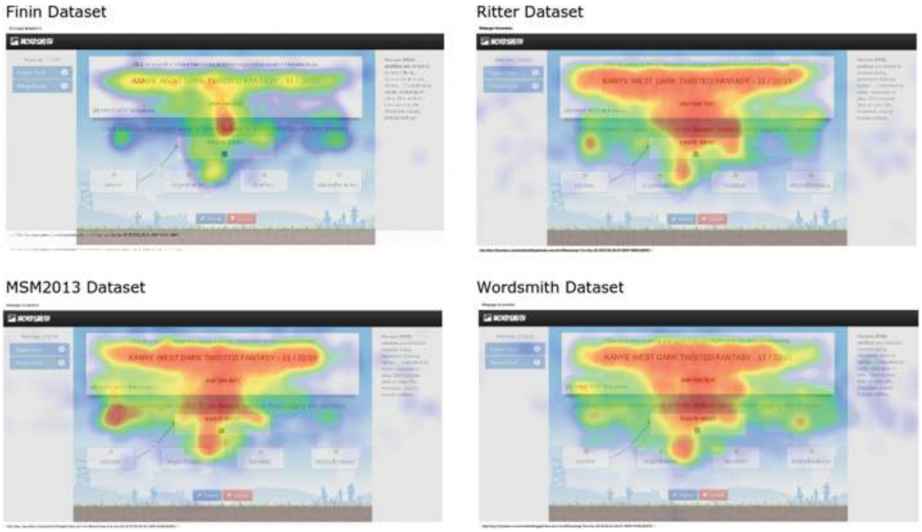
Our own *Wordsmith dataset* achieved the highest precision and recall values in identifying people and places. Again, crowd workers had trouble classifying entities as MISC and significant noise hindered the annotation of ORG instances. A number of ORG entities were misidentified as PER and an equally 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).

Table 2. *Experiment results* - Named entity recognition on the four datasets.

Ritter dataset							
Worker annotations				Confusion matrix (vs gold)			
Entity type	Precision	Recall	F1 score	PER	ORG	LOC	MISC
Person	42.93	**69.19**	52.98	765	7	26	20
Organisation	28.75	39.57	33.30	10	140	62	88
Location	**67.06**	50.07	**57.33**	9	9	751	22
Miscellaneous	20.04	20.23	20.13	15	46	29	217
Finin dataset							
Worker annotations				Confusion matrix (vs gold)			
Entity type	Precision	Recall	F1 score	PER	ORG	LOC	MISC
Person	**68.42**	58.96	**63.34**	78	1	7	-
Organisation	50.94	27.84	36.00	1	27	5	-
Location	66.14	**60.71**	63.31	1	4	84	-
Miscellaneous	-	-	-	-	-	-	-
MSM2013 dataset							
Worker annotations				Confusion matrix (vs gold)			
Entity type	Precision	Recall	F1 score	PER	ORG	LOC	MISC
Person	**87.21**	**86.61**	**86.91**	3,828	25	8	7
Organisation	43.27	38.77	40.90	16	299	13	28
Location	60.57	67.29	63.75	13	21	321	5
Miscellaneous	10.44	29.11	15.37	12	82	5	91
Wordsmith dataset							
Worker annotations				Confusion matrix (vs gold)			
Entity type	Precision	Recall	F1 score	PER	ORG	LOC	MISC
Person	**79.23**	71.41	**75.12**	5,230	34	29	32
Organisation	61.07	53.46	57.01	93	811	30	46
Location	72.01	**72.91**	71.26	25	58	1,078	8
Miscellaneous	27.07	47.43	34.47	50	113	12	718

5.2 Analysis of Behavioral Features of Crowd Workers

The results on the skipped true-positive tweets are presented in Fig. 4. 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 experiments, people tend to avoid recognizing organisations, but were more keen in identifying locations. Disambiguating between the two remained challenging across all datasets as evidenced in the confusion matrices in Table 2. 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



Finin Dataset

Ritter Dataset

MSM2013 Dataset

Wordsmith Dataset

Fig. 3. Wordsmith Heatmaps across the 4 datasets

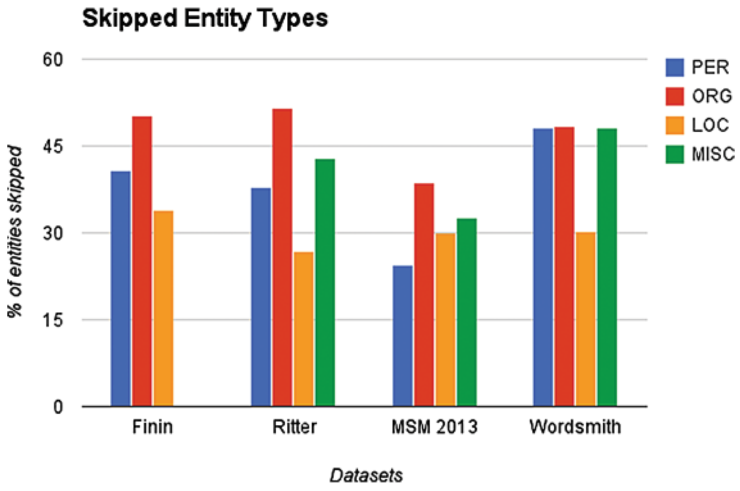


Fig. 4. Skipped tweets

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 MSM2013 dataset, person entities were least skipped due to the features of the dataset discussed earlier (e.g., clear text definition, consistent capitalisation etc.). In the Wordsmith dataset, however, PER, ORG, and MISC entity tweets were skipped with equal

Table 3. *Experiment results* - Skipped true-positive tweets

Skipped tweets				
Dataset	Skipped		Annotated	
	No. entities	Tweet length	No. entities	Tweet length
Finin	1.56	101.39	1.33	94.82
Ritter	1.42	113.05	1.35	104.22
MSM 2013	1.49	98.74	1.30	97.11
Wordsmith	1.62	102.22	1.39	97.84

Table 4. *Experiment results* - Average accurate annotation time

Average accurate annotation time (seconds)				
Dataset	PER	ORG	LOC	MISC
Finin	9.54	12.15	8.91	-
Ritter	9.69	10.05	9.35	10.88
MSM 2013	9.54	10.77	8.70	10.35
Wordsmith	8.06	8.50	9.56	9.48

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.

Table 3 gives further insight into those microtasks that remained unsolved. The results show for each dataset the average number of entities present in the skipped and un-skipped tweets, alongside the average length for both categories. We note that on average, workers preferred to avoid relatively long posts and those containing more entities. The tweet length was least significant in the MSM2013 experiment, once again due to the comparatively well-formed nature of the dataset and the least standard deviation in the tweet lengths. This 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.

Table 4 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 show the shortest time needed corresponds to 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. As discussed earlier, this can be attributed also to entities arising from *@usernames* and *#hashtags*. The other datasets either exclude this or do not annotate it in their gold standards.

Figure 3 shows the result of our datapoint captures via heatmaps. The results show 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 Fig. 1, we can reconcile the mouse movements to point predominantly to “PERSON” and “LOCATION” entities in proportions which are consistent with the individual numbers presented in Table 2.

6 Discussion and Conclusion

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].

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 increase 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 [6,11]. The only exception for this is the PER entity type, which was the most accurate result for the MSM dataset.

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 class of entities which were not in the gold standard but were annotated by the crowd are what we call *implicitly named entities*. Examples such as *hair salon*, *last stop*, *in store*, or *bus stop* show 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 observe 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.

Wide search, but centred spot. Our heatmap analysis shows 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 is higher than for the gold standard. This might mean that people are more likely to miss out on annotating entities on the left and right edges of the interface. A resolution could be to centralize the textbox and make it less wide hence constraining the worker's field of vision as opposed to [10] where workers were required to observe vertically to target entities. We cannot fully substantiate this claim yet and reserve this for further work due to the responsive nature of the interface which would have presented the annotation text slightly different on varying screen resolutions and with screen resizings.

Concluding remarks and future work. In this paper we have experimented with a novel approach to finding entities within micropost datasets using crowdsourced methods. Our experiments, conducted on four different micropost datasets, have 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 devising automated approaches to determining when best to select human or machine capabilities, and also examining *implicitly named entities* in order to develop methods to identify and derive message-related context and meaning.

References

1. Aroyo, L., Welty, C.: Crowd truth: harnessing disagreement in crowdsourcing a relation extraction gold standard. In: WebSci2013. ACM (2013)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) ISWC/ASWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007)
3. Basave, A.E.C., Varga, A., Rowe, M., Stankovic, M., Dadzie, A.: Making sense of microposts (# msm2013) concept extraction challenge. In: # MSM, pp. 1–15 (2013)
4. Braunschweig, K., Thiele, M., Eberius, J., Lehner, W.: Enhancing named entity extraction by effectively incorporating the crowd. In: BTW Workshops 2013, pp. 181–195 (2013)
5. Demartini, G., Difallah, D.E., Cudré-Mauroux, P.: Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In: Proceedings of the 21st International Conference on World Wide Web, pp. 469–478. ACM (2012)
6. Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., Bontcheva, K.: Analysis of named entity recognition and linking for tweets. Inf. Process. Manag. **51**(2), 32–49 (2015)
7. Difallah, D.E., Demartini, G., Cudré-Mauroux, P.: Mechanical cheat: spamming schemes and adversarial techniques on crowdsourcing platforms. In: CrowdSearch, pp. 26–30 (2012)
8. Feyisetan, O., Simperl, E., Tinati, R., Luczak-Roesch, M., Shadbolt, N.: Quick-and-clean extraction of linked data entities from microblogs. In: Proceedings of the 10th International Conference on Semantic Systems, SEM 2014, pp. 5–12. ACM (2014)
9. Feyisetan, O., Simperl, E., Van Kleek, M.: 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)

10. 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. Association for Computational Linguistics (2010)
11. Fromreide, H., Hovy, D., Søgaard, A.: Crowdsourcing and annotating NER for Twitter #drift. European language resources distribution agency (2014)
12. 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. Association for Computational Linguistics (2010)
13. Liu, X., Zhang, S., Wei, F., Zhou, M.: Recognizing named entities in tweets. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 359–367. ACL (2011)
14. Marrero, M., Sanchez-Cuadrado, S., Lara, J.M., Andreadakis, G.: Evaluation of named entity extraction systems. Adv. Comput. Linguist. Res. Comput. Sci. **41**, 47–58 (2009)
15. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1), 3–26 (2007)
16. Ritter, A., Clark, S., Etzioni, O., et al.: Named entity recognition in tweets: an experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1524–1534. Association for Computational Linguistics (2011)
17. Rizzo, G., Troncy, R.: Nerd: evaluating named entity recognition tools in the web of data (2011)
18. Sabou, M., Bontcheva, K., Derczynski, L., Scharl, A.: Corpus annotation through crowdsourcing: Towards best practice guidelines. In: Proceedings of LREC (2014)
19. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: 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, pp. 254–263. Association for Computational Linguistics (2008)
20. Usbeck, R., Ngonga Ngomo, A.-C., Röder, M., Gerber, D., Coelho, S.A., Auer, S., Both, A.: AGDISTIS - graph-based disambiguation of named entities using linked data. In: Mika, P., et al. (eds.) ISWC 2014, Part I. LNCS, vol. 8796, pp. 457–471. Springer, Heidelberg (2014)
21. von Ahn, L., Dabbish, L.: Designing games with a purpose. Commun. ACM **51**(8), 58–67 (2008)
22. 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)
23. Yetisgen-Yildiz, M., Solti, I., Xia, F., Halgrim, S.R.: 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, pp. 180–183. Association for Computational Linguistics (2010)
24. Yuen, M., King, I., Leung, K.: A survey of crowdsourcing systems. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (Passat) and 2011 IEEE Third International Conference on Social Computing (SocialCom), pp. 766–773. IEEE (2011)