Semantic Web 0 (0) 1–26
IOS Press

An Extended Study of Content and Crowdsourcing-related Performance Factors in Named Entity Annotation

Editor(s): Marta Sabou, Technische Universität Vienna, Austria; Lora Aroyo, Vrije Universiteit Amsterdam, Netherlands; Kalina Bontcheva, University of Sheffield, UK; Alessandro Bozzon, Technische Universiteit Delft, Netherlands
Solicited review(s): Leon Derczynski, The University of Sheffield, UK; Two anonymous reviewers

Oluwaseyi Feyisetan [a,*], Elena Simperl [b], Markus Luczak-Roesch [c], Ramine Tinati [b], and Nigel Shadbolt [d]

[a] *University of Southampton, Southampton, UK*
E-mail: oof1v13@soton.ac.uk
[b] *University of Southampton, UK*
E-mail: {e.simperl,r.tinati}@soton.ac.uk
[c] *Victoria University of Wellington, New Zealand*
E-mail: markus.luczak-roesch@vuw.ac.nz
[d] *University of Oxford, UK*
E-mail: nigel.shadbolt@jesus.ox.ac.uk

Abstract. Hybrid annotation techniques have emerged as a promising approach to carry out named entity recognition on noisy microposts. In this paper, we identify a set of content and crowdsourcing-related features (number and type of entities in a post, average length and sentiment of tweets, composition of skipped tweets, average time spent to complete the tasks, and interaction with the user interface) and analyse their impact on correct and incorrect human annotations. We then carried out further studies on the impact of extended annotation instructions and disambiguation guidelines on the factors listed above. This was all done using CrowdFlower and a simple, custom built gamified NER tool on three datasets from related literature and a fourth newly annotated corpus. Our findings show that crowd workers correctly annotate shorter tweets with fewer entities, while they skip (or wrongly annotate) longer tweets with more entities. Workers are also adept at recognising people and locations, while they have difficulties in identifying organisations and miscellaneous entities which they skip (or wrongly annotate). Finally, detailed guidelines do not necessarily lead to improved annotation quality. 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 to automatic tools, crowdsourced workers and nichesourced experts. Experimental results are published as JSON-LD for further use.

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

1. Introduction

Harnessing the rapid increase in the generation of data has led to advances in the Semantic Web and the Web of Data vision [3]. A first step in making sense of the data necessitates information extraction and annotation of datasets. This has led to the availability of training datasets for Natural Language Processing algorithms from research such as ACE [18], MUC [10] and CoNLL [42]. An important task in this context is the identification of named entities - the people, places, organisations, and dates referred to in text documents -

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

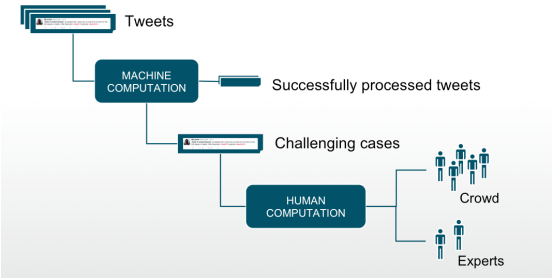
1570-0844/0-1900/$27.50 © 0 – IOS Press and the authors. All rights reserved

and their mapping to Linked Data URIs [47]. 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 [29,33]. It has been less successful so far in processing social media content such as microblogs, known for its compact, idiosyncratic style [15]. Human computation and crowdsourcing offer an effective way to tackle these limitations [45], alongside increasingly sophisticated algorithms capitalising on the availability of huge data samples and open knowledge bases such as DBpedia and Freebase [38].

Advances in natural language processing has led to an understanding of textual structure which can be easily processed by computers (e.g.,, well formed news-wire articles with sufficient disambiguation context). Essentially, hybrid workflows have therefore led to pipelines which first selects text for machine annotation, passing the residue to the crowd (such as the approach by [14]). These hybrid approaches to NER (named entity recognition) [15] that seamlessly bring together human and computational intelligence are however 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 [7,40] and CrowdDB attempts crowd powered query engines [46] - 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., [17,45,53]), 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.

In our work, we posit that just as certain textual features (such as proper syntax and sufficient context) make them amenable to automatic NER, certain features also lead to higher quality named entity annotation by crowd workers. This could lead to the design of more advanced workflows as shown in Figure 1 where the initial processing divides tweets between automatic tools and the crowd, and subsequently between the crowd and experts. This paper offers an in-depth study of the factors which influence the performance of the crowd in hybrid NER approaches for microposts. We categorise these features in 2 broad classes: 1) content features inherent in the tweets such as - number of entities, types of entities, length of the tweet and the tweet sentiment; and 2) crowdsourcing features and factors observed during annotation such as - skipped true-positive posts, average time spent to complete the tasks, accuracy of the answers and the worker interaction with the user interface. We analyse the impact of these features on the accuracy of the results, the timeliness of their delivery and their distribution in correct and incorrect annotations. In order to fully understand these factors, we also studied the importance of annotation guidelines vis-à-vis the debate on the role of detailed guidelines as a means of improving human annotation [2].

Fig. 1. Hybrid Workflow



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) [49] called Wordsmith.[1] An analysis of the overarching results reveal that detailed guidelines do not necessarily lead to higher quality annotations. The presence of additional disambiguating information however leads to specific annotation improvements such as annotating #hashtags and @mentions. Further analysis of the results show 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.

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 hu-

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

man 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

2.1. Crowdsourced NER

Several approaches have been applied to build tools for entity extraction, using rules, machine learning, or both [27]. An analysis of the state of the art in named entity recognition and linking on microposts is available in [15]. 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 [19,37].

Crowdsourcing has been previously used to annotate named entities in micropost data [21]. 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. 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 [21] was carried out and reported in [22]. 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 [15].

A similar approach has been used to carry out NER tasks on other types of data. Lawson et al [26] 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 [52]. 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. [50] presented a hybrid approach where expert annotators identified the presence of entities while crowd workers assigned entity types to the labels. [14] 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. Using hybrid systems to offer crowd based query processing has also been studied by [46]. Their work leveraged on the crowd to improve recall scores in open ended questions and how a mixed crowd can help converge on an accurate answer. Other examples of similar systems are [8] and [40]. [40] 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 [51].

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.

2.2. Crowd worker performance

One of the earlier works focusing on utilising the crowd for annotation tasks was by [45] where they used a pre-computed gold standard to improve annotator quality. Several other approaches has been presented to improve the quality of task output by crowd workers. These include using detailed annotation guidelines; engaging multiple annotators [26] and relying on results with high inter-annotator agreements. A set of guidelines for corpus annotation, distilled from existing literature was presented by [7]. Of note are the sections on *in-task quality*, *contributor evaluation* and *aggregation* where various approaches such as the use of gold standards, majority voting, active learning and average reliability are mapped to their adoption in literature. The role of uncertainty arising from worker annotation was addressed by [35] by looking at inter-annotator agreement loss. Also of importance in crowdsourced annotation is the role of worker diversity [46] which improves recall by unearthing patterns which could not be seen by a homogeneous set of limited experts. Further factors also affect worker quality beyond the presence of a diverse crowd. Some extrinsic factors affecting annotation quality were presented by [12].

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 two 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 evaluate crowd worker preferences for NER tasks.

RQ2.1. Can we evaluate crowd workers preferences for certain types of tasks by observing and measuring (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. The term GWAP here is used lightly - as we did not design Wordsmith within the context of this study to include features which occur in traditional games (or gamified systems) such as points, badges and leaderboards. Wordsmith however supports more bespoke functions which could not be easily achieved by using CrowdFlower. We describe the platform in more detail in Section 6

4.1. Research data

We took three datasets from related literature, which were also reviewed by [15]. 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 [37] 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 [15]). 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 [21] 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 [19]. 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 [38] 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 [22]. Furthermore, all four used different entity classification schemes, which we normalized using the mappings from [15]. Table 1 characterizes the data sets along the features we hypothesize might influence crowdsourcing effectivity. We further refer the interested reader to an even more recent dataset – [16] by the authors of [15].

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 $\$0.05$ to annotate at least 10 tweets with no bonus incentive. We will discuss these choices in 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

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

annotation instruction; (ii) entity type disambiguation instruction and (iii) an updated interface which presented 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 [23,41]. We hypothesized that particularly polarised tweets might have an effect on the entity annotation [32]. 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. AlchemyAPI was also used to carry out sentiment analysis on movie reviews from IMDb by [44]. Their results presented AlchemyAPI with an F1 score of 77.78% on a dataset of $1,000$ reviews.

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. We also aggregated the performance of users in order to identify a number of distinguishing behavioural characteristics related to 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 [38] to normalise these definitions even though the results were slightly different from the entity mappings adopted by [15]. Our mappings (see Table 2) 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.

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

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 *Walmart* in context becomes difficult, even for expert annotators. This extends to other classes such as museums, restaurants, universities and shopping malls. Our disambiguation approach is presented in Table 3.

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, the 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.

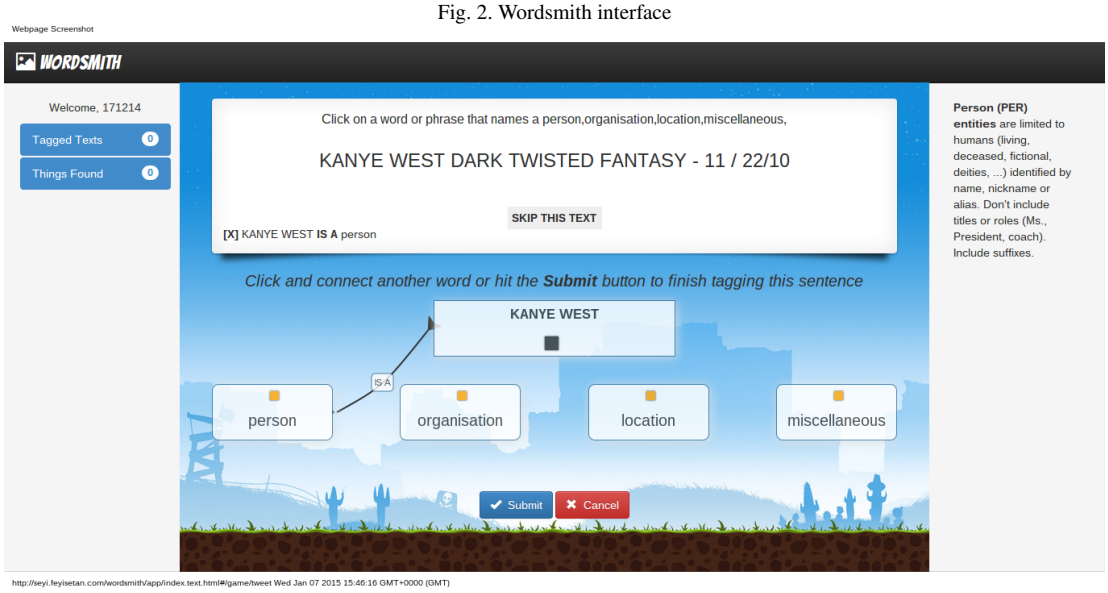
6. Crowdsourcing approach

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.

Fig. 2. Wordsmith interface

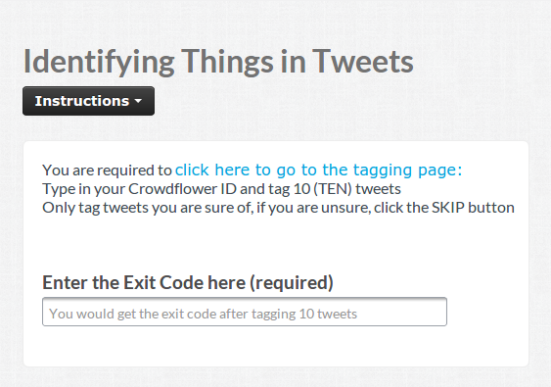


The main interface of Wordsmith is shown in Figure 2. It consists of three sections. The annotation area is at the center of the screen with sidebars for additional information. The tweet under consideration is presented at the top of the screen with each text token presented as a highlight-able span. 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 [21] and span selections by [8,26,50]. We opted for a click-and-drag approach in order to fit all the annotation components on the screen (as opposed to [21]) and to cut down the extra type verification step by [8]. 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 3, 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 [20]. 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 [26,52], we did not use any bonuses. The annotations carried out in [26] 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 [52] 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 an-

Fig. 3. CrowdFlower interface



notation 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 [20]; 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. We display the second tweet at a fixed point in order to simplify our analysis and remove bias arising from workers viewing the tweet at random intervals.

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 [21] for person (PER), organisation (ORG) and location (LOC) entity types. We also included a fourth miscellaneous (MISC) type, based on the guidelines from [37].

In computational linguistics, annotation guidelines present arbitrary and often debatable decisions [35] as seen from the varying choices in our experiment datasets. The decision to annotate (or not to) *#hashtags*, *@mentions* and MISC types represent the beginning of choices which extends to guidelines on specific entity types. Some authors have argued that more detailed guidelines do not improve annotation quality [2]; while some others skip the guidelines altogether when dealing with experts [35]. The latter category rely on the experts to make adhoc consensual judgements amongst themselves to address hard cases.

In our study, we experimented with 2 guideline conditions to observe the results of varying the amount of annotation guidelines.

Experiment condition 1
Instructions were presented at the start of the CrowdFlower job 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 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 [34] and is seen as a way to judge the reliability of annotated data [36]. 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 [11] have been used to calculate inter-rater agreement.

The inter-annotator agreement (or degree of disagreement) can also serve as a measure of the difficulty of the task - and can draw light unto 'hard cases' which might require further attention [1,35]. Annotator disagreement is not limited to crowd workers only but extends to experts also. The authors of [1] argue that inter-annotator disagreement is *not noise, but signal*; and, [35] incorporates it in the loss function of a structured learned for POS tagging and NER.

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

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

7. Results

7.1. Overview

Overview of Annotations

Table 5 gives an overview into how workers performed at the tweet level across the various datasets. The results suggests consistently that workers correctly annotate tweets with fewer entities. This result was consistent across the four datasets. We did not see any strong connection between the length of the tweet and the likelihood of it being annotated correctly or incorrectly as the differences were not significant. The length of the tweet however determines the whether the tweet would be selected for annotation or not – and we discuss this in detail in a later section.

Correct Annotations

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

	Condition 1: Worker annotations			Condition 2: Worker annotations		
Entity type	Precision	Recall	F1 score	Precision	Recall	F1 score
Finin dataset						
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						
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						
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						
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.

Correct and Incorrect Annotations				
Dataset	**Correct**		**Incorrect**	
	Num of Entities	Tweet length	Num of entities	Tweet length
Finin	1.17	91.63	1.48	92.53
Ritter	1.24	106.02	1.61	99.02
MSM	1.19	98.95	1.81	97.02
Wordsmith	1.38	97.88	1.70	96.10

Table 5

Experiment results - Correct and Incorrect Annotations

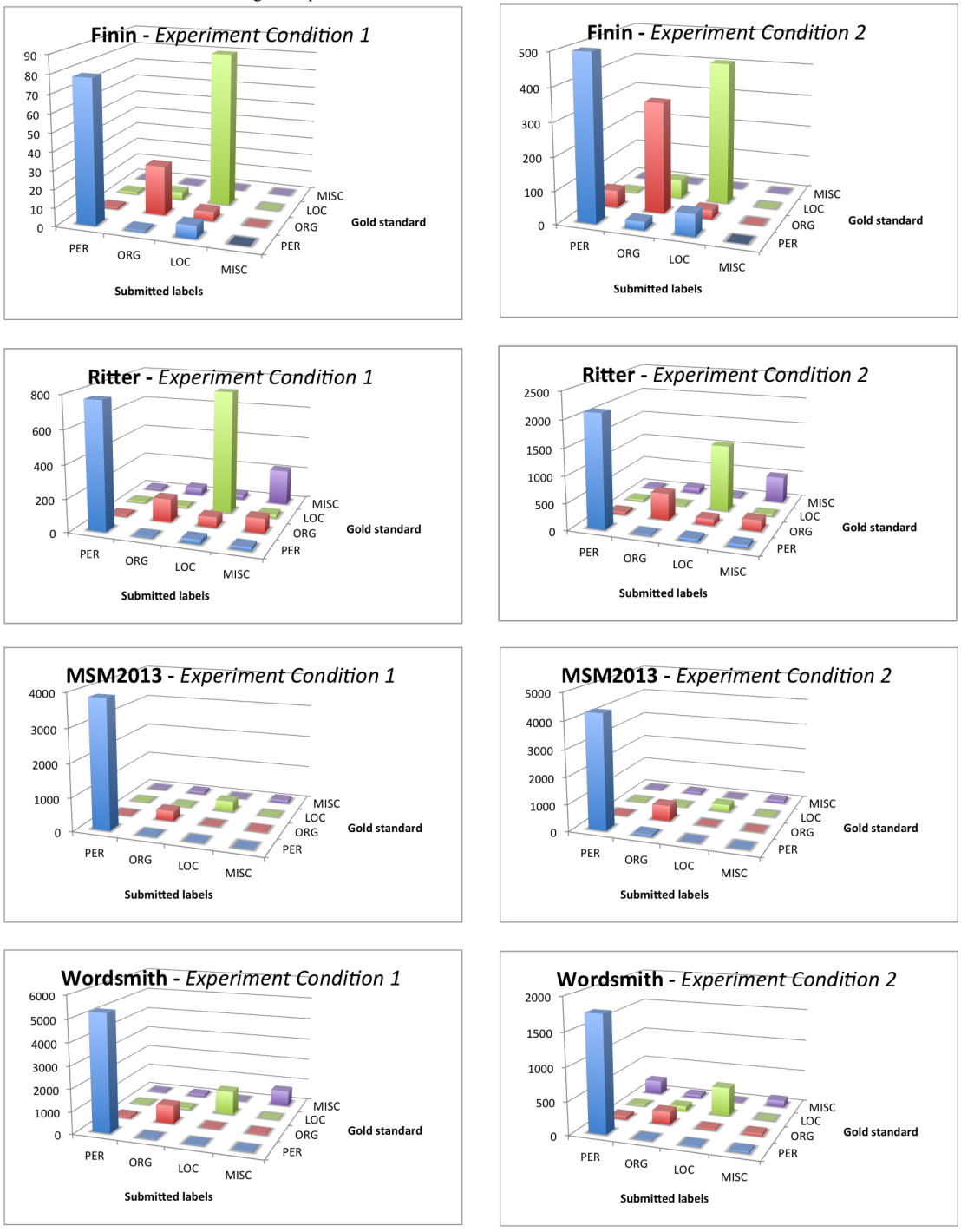
from C1 to C2. This is in spite of the additional annotation guidelines presented in C2. This result is in line with *Myth 3* presented by [2] which states that detailed guidelines do not always yield better annotation quality. The results show highest precision scores in identifying PER entities. The 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 MSM2013 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 further observe from Table 4 that the precision and recall scores vary quite markedly across the different datasets. There are a number of factors that contribute to this, most of which would be discussed in the ensuing sections and summarised in Section 7.2.1. Particularly, we point out the low scores for the Ritter dataset. It is important to note that the dataset does not annotate Twitter *@usernames* and *#hashtags* (given that many *@usernames* are labelled as people and organisations). It can also be attributed to the annotation schema as noted by [15] – for example, Ritter assigns the same entity type *musicartist* to single musicians and group bands.

Incorrect Annotations

Figure 5 illustrates the entity types which were wrongly annotated by workers. Across all the datasets, we observe that the ORG and MISC entity types were consistently wrongly annotated. This was the case across the four datasets. This suggests that workers had the greatest difficulties in either identifying these entity types, or were mis-assigning them to other entity types. We computed a confusion matrix (presented in Figure 4,

Fig. 4. *Experiment results* - Confusion Matrix on the four datasets

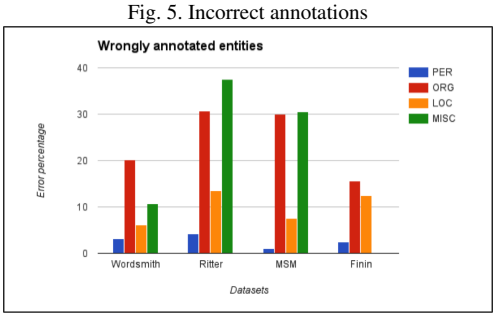


and discussed hereafter) to have a clearer insight into what entity types were wrongly annotated, and how they were wrongly annotated.

Mismatched Annotations

We included a confusion matrix in Figure 4, highlighting the entity mismatching types e.g., assigning *Cleveland* as location when it refers to the basketball team.

Fig. 5. Incorrect annotations



The matrix layout is inspired by the confusion matrix diagrams illustrated in [48] which utilises a 3 dimensional columnar graph to draw attention to the relationship between the worker submitted labels and the gold standards. The height of each bar corresponds to the number of submissions for that entity type. The *x-axis* represents the worker submissions, the *z-axis* identifies the gold standard while the diagonals indicate the intersection of the worker annotations and the gold labels.

The results suggest 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). The typical confusion of the ORG and LOC types is a case of metonymy where these entities have to be especially handled in context [30]. This is seen where an organisation is associated with its location e.g., *Wall Street* and *Hollywood*. This phenomenon occurred in both experiment conditions 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.

In comparison with the precision and recall scores earlier presented in Table 4, the numbers presented in the confusion matrix of Figure 4 represent the total annotation count by the participating workers as opposed to an aggregate. This excludes malicious workers, spammers and other outliers that might distort the overall results.

Skipped Tweets: Tweet Overview

Our guidelines encouraged workers to skip tweets for which they could not give confident annotations. Table 6 gives further insight into the dynamics of skipped tweets. The table presents, for C1 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.

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 6

Experiment results - Skipped true-positive tweets

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.

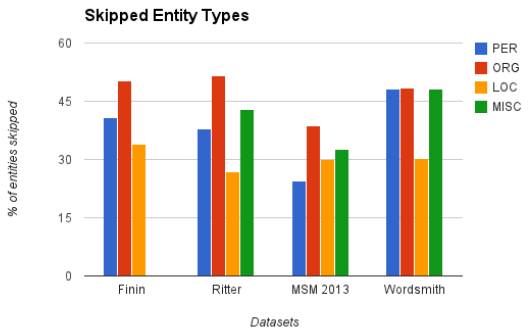
We do not report a high level metric on the number of tweets skipped as this might have been misleading. For example, given 10 tweets annotated by 3 workers, the tweets skipped by each worker might

have been annotated by another. We therefore present fine grained results on the distribution of entity types present in tweets skipped by individual workers and the tweet sentiment. We also report aggregate findings on the average number of entities present in, and the average length of skipped tweets.

Skipped Tweets: Entity Types

More results on the skipped true-positive tweets are presented in Figure 6 and Table 7. 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 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.). The entity types in the Wordsmith dataset (apart from the LOC type) were all skipped with equal likelihoods.

Fig. 6. *Skipped Tweets*: Entity Types in Skipped Tweets



We posit this to be as a result of two factors: our uniform sampling method which did not bias the presence of a single entity type (e.g., as in the MSM2013 dataset) and increased use of *@mentions* and *#hashtags* in the dataset. This result is also in line with those presented in Figure 4 that ORG was the most misidentified entity type. This result was consistent across both experiment conditions with crowd workers still skipping tweets with organisation entities when more instructions were given on how to disambiguate them.

Skipped Tweets: Sentiment Analysis

Table 8 summarises the sentiment distribution of positive, negative and neutral tweets in the different datasets. The results present the Finin, Ritter and MSM2013 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 9 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.

Annotation Time: On Correct Annotations

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 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. We posit that the extended annotator guidelines contributed to the decrease in annotation time. As this was the variable in this condition, our hypothesis is that a more detailed level of annotation guidelines leads to an anchored and increased confidence amongst the annotators. This in turn leads to mechanistic annotations - i.e. spotting a text and annotating it according to the guideline without discerning the relevant context. This can explain for the increase in speed which did not necessarily result in an increase in annotation quality.

Interface and Heatmaps

Figure 7 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

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 7

Skipped Tweets - Skipped tweets containing entities

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 8

Sentiment Analysis - General distribution

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)

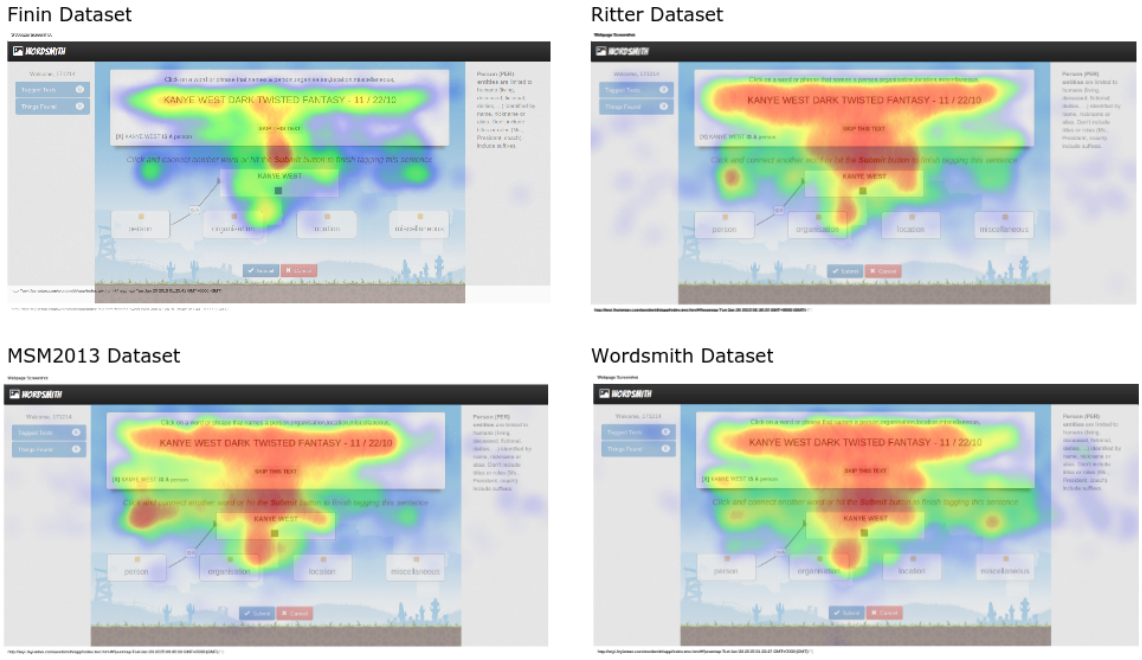
Table 9

Skipped Tweets - Sentiment analysis distribution of skipped tweets

which indicate which entity types were tagged. From a larger image of the interface in Figure 2, we can reconcile the mouse movements to point predominantly to PER and LOC entities in proportions which are consistent with the individual numbers presented 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

Fig. 7. Wordsmith Heatmaps across the 4 datasets



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

standards vary from the 14th character in the Wordsmith dataset to the 35th character in the MSM2013 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.

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

Inter-Annotator Agreement

Table 12 summarises the average inter-annotator agreement scores across the four datasets. Based on our design choices, workers were allowed to skip tweets which they could not confidently annotate. Workers were required to annotate at least 10 tweets and each tweet was annotated by at least 3 annotators. The results presented here represents the inter-annotator agreement on tweets which were annotated by 3, 4, 5 and 6 workers each. At a high level, the results show that agreement begins to break down as consensus is required amongst more workers. This is not surprising as a base agreement between 2 out of 3 workers is equivalent to 66.67%. Drawing workers out of the same distribution on a tweet annotated by 4 workers yields a lower score of 50%. This interprets the decline

in inter-annotator agreement scores as more workers annotated the same tweet.

Dataset	Number of Annotators			
	3	4	5	6
Finin	**62.40**	53.84	48.39	49.47
Ritter	**62.28**	52.84	47.11	39.03
MSM	**83.47**	83.08	79.80	77.86
Wordsmith	**60.28**	57.03	50.16	41.90

Table 12

Experiment results - Average Inter-Annotator Agreement

The inter-annotator agreement scores were clearly highest in the MSM 2013 dataset (83.47%). This can be attributed to the relative homogeneity of the dataset and the presence of a large number of easily identifiable PER entities. The other 3 datasets had similar scores with an average inter-annotator agreement of 61.65% and a standard deviation of 1.19.

Entity Inter-Annotator Agreement				
Dataset	PER	ORG	LOC	MISC
Finin	51.68	23.07	47.95	18.27
Ritter	68.05	13.67	34.14	14.69
MSM	86.95	13.20	33.72	10.62
Wordsmith	70.68	13.47	40.38	11.42

Table 13

Experiment results - Entity Level Inter-Annotator Agreement

In Table 13, we drill further into the inter-annotator agreement on the entity level. The results presented in this table were based on the results of 3 annotators per tweet (extrapolated from the first column in the results within Table 12). The results are in line with earlier results presented i.e. workers are better at identifying PER and LOC entities (as these entity types receive the highest scores), and have greater difficulties with ORG and MISC entities.

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 MSM2013 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 MSM2013 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 precision scores are also from the Wordsmith dataset in identifying organisations which had a boost of about 30%. Precision scores in the MSM2013 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 MSM2013 dataset.

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 [15]). 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. 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).

Disambiguating between ORG and LOC types remained challenging across all datasets as evidenced in the confusion matrices in Figure 4. 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 present these results in light of tweets which were annotated correctly, incorrectly and skipped tweets. 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 Table 6 we see that the number of entities in a tweet affect the likelihood of annotation by a worker i.e., regardless of whether the annotations are accurate or not, a tweet with fewer entities was more likely to be selected. 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. Longer tweets were therefore more frequently skipped by workers.

The results in Table 5 give further insight into the role of the number of entities in correctly and incorrectly annotated tweets. The results show consistently across the 4 datasets that once a tweet has been selected for annotation, it is more likely to be annotated correctly and completely if it has fewer entities, while tweets with more entities were wrongly annotated. In summary, skipped tweets (more entities), incorrect tweets (less than skipped tweets), correct tweets (even less than both).

Entities types

Figure 6 and Table 7 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 even though the ORG type was not the most common 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 MSM2013 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. Overall, 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 MSM2013 dataset had a high number of PER entities, however, these were not skipped as the tweets were from well structured texts e.g., quotes with the author attribution at the end.

Micropost text length
The resuslts presented in Table 5 and Table 6 suggest that the tweet length was a factor in determining whether it was selected for annotation or not (since workers were free to select what tweet they annotated). However, after the tweet has been selected, there was no strong connection between the length of the tweet and the annotation accuracy. The standard deviation of the datasets was 5.65 characters, however, the standard deviation of tweets selected for annotation was 3.41 characters. As a result, at the selection stage, the tweet length played a role in the likelihood of a worker deciding to annotate, however, the length did not further matter as most of the tweets were of similar lengths.

Table 6 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 above 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 result was consistent in all datasets asides the Ritter dataset which had an overall set of longer tweets. From this we observe that regardless of the dataset (such as the MSM2013 dataset with an average length of 88 characters), workers would be willing to annotate up to a certain threshold before they start skipping.

These results might not be unconnected with the user interface design. Revisiting our interface in Figure 3 gives an insight into how the tweets appear in the annotation interface. Shorter tweets would fit squarely in the task box with minimal text wrapping. This layout is similar to [6] in that the GATE annotation tool also lays out the tweet horizontally (for workers to annotate from left to right) unlike [21] which lays the tweet vertically (for workers to annotate from top to bottom). Interpreting this further in the light of the results in Table 11 might suggest that workers were annotating entities immediately within their field of vision since they consistently started annotating at a given point across all the datasets.

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.

7.2.3. 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 6, 7, and 9 give insights into the skipped tweets. The results show that across the datasets, the number of entities and the length of the tweet were two factors that contributed to the likelihood of a skipped tweet. Table 7 further highlights the role entity types play on workers choosing to annotate a tweet or not. At this time we cannot present conclusive remarks on the effect of the tweet sentiment on a workers probability of annotating it.

Apart from these high level features such as the number and type of entities, and the micropost length, we also discovered some other latent features which might contribute to workers skipping tweet. For example, a closer look at the Wordsmith dataset (which was the most recent corpus) revealed that workers skipped the various entity types with almost equal likelihoods. We reported this as being tied to an increase in the use of *#hashtags* and *@mentions*. Furthermore, the corpus contained #hashtags referencing events such as the #WorldCup2014 and #LondonFashionWeek which created annotation ambiguity. In the second experi-

ment condition C2, workers spent less time annotating and skipped fewer entities due to the availability of detailed guidelines. As noted earlier, this helped workers disambiguate some entity types (e.g. handling entities from #hashtags), however, it did not result in an overall improvement in annotation quality.

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. Figure 4 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 7. 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 labelling 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 MSM2013 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.

Implicitly named entities
In our investigation we paid special attention to those entities that were annotated by the crowd but that were not covered by the gold standard. As a result of a manual inspection of these cases one particular category of entities stands out, which we call *implicitly named entities*. By that term we mean those entities that were represented in the text by a proxy phrase that – if the user's contextual assumptions are known – one can infer an actual named entity. A particular example for this is the annotated phrase *'last stop'*, which, if one would know the place, direction and means of transportation to contextualize the annotation, could be resolved to one explicit stop or station.

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 other 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 preprocessing them [43]. One option would be to try out a multi-step workflow (such as the hybrid workflow proposed by [39]) in which entity types that 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 [45], 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 MSM2013 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 [15,22]. The only exception for this is the PER entity type, which was the most accurate result for the MSM2013 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 [9,31].

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 recognition 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: (i) unnamed entities, (ii) partial entities, (iii) overlapping entities, and (iv) hashtags. It is noteworthy that we observed an additional 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 or a place in a city that would not fall into the category of a typical point of interest). This is an important finding for hybrid entity extraction pipelines, which can benefit from the capability to generate new URIs for yet publicly unknown entities. This can play an important role in modern (data) journalism [28] and complements the findings about the entity annotation behavior of technical non-experts on longer texts presented in [24] and [25].

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 [21] 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 on the average spent less time annotating entities; (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 [20] 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 [13]. 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.

Revisiting the role of experts Some of the results presented here might ferment questions on the usefulness of the crowd in carrying out high quality named entity recognition on noisy microposts. Indeed, the crowd is but one step in the workflow required to achieve the Web of Data vision and understanding how to harness their unique capabilities is of utmost importance. Automatic annotation processes have continued to improve and this has been in part due to the availability of pre-annotated corpora - carried out by experts and the crowd. We believe our work would form one of the missing components in addressing the design of more advanced workflows which could necessitate the reintroduction of experts into the loop - fitting in to disambiguate where the crowd falls short.

In addition, the crowd helps to shed further light into what might have been overlooked by a trained set of experts, opening up potentials out of scope of predefined research questions. For example, in our case, the potentials of implicit entities could help in the design of conversational AI assistants which could resolve *last stop*, *in store*, or *bus stop* based on context.

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] Lora Aroyo and Chris Welty. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In Hugh Davis, Harry Halpin, Alex Pentland, Mark Bernstein, Lada Adamic, Harith Alani, Alexandre Monnin, and Richard Rogers, editors, *Proceedings of the 3rd Annual ACM Web Science Conference, 2013, Paris, France, WebSciâĂŸ13*. 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, 2015. DOI https://doi.org/10.1609/aimag.v36i1.2564.

[3] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. DBpedia: A nucleus for a web of open data. In Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007.*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer, 2007. DOI https://doi.org/10.1007/978-3-540-76298-0_52.

[4] Amparo Elizabeth Cano Basave, Andrea Varga, Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie. Making Sense of Microposts (#MSM2013) concept extraction challenge. In Amparo Elizabeth Cano, Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie, editors, *Proceedings of the Concept Extraction Challenge at the Workshop on 'Making Sense of Microposts', Rio de Janeiro, Brazil, May 13, 2013*, volume 1019 of *CEUR Workshop Proceedings*, pages 1–15. CEUR-WS.org, 2013. URL http://ceur-ws.org/Vol-1019/msm2013-challenge-report.pdf.

[5] Plaban Kumar Bhowmick, Anupam Basu, and Pabitra Mitra. An agreement measure for determining inter-annotator reliability of human judgements on affective text. In Ron Artstein, Gemma Boleda, Frank Keller, and Sabine Schulte im Walde, editors, *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics, 23 August 2008, Manchester, UK*, pages 58–65. Coling 2008 Organization Committee, 2008. URL http://www.aclweb.org/anthology/W08-1209.

[6] Kalina Bontcheva, Ian Roberts, Leon Derczynski, and Dominic Paul Rout. The GATE crowdsourcing plugin: Crowdsourcing annotated corpora made easy. In Gosse Bouma and Yannick Parmentier, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 97–100. Association for Computational Linguistics, 2014. URL http://aclweb.org/anthology/E/E14/E14-2025.pdf.

[7] Kalina Bontcheva, Leon Derczynski, and Ian Roberts. Crowdsourcing named entity recognition and entity linking corpora. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*. Springer, 2017. To appear.

[8] Katrin Braunschweig, Maik Thiele, Julian Eberius, and Wolfgang Lehner. Enhancing named entity extraction by effectively incorporating the crowd. In Gunter Saake, Andreas Henrich, Wolfgang Lehner, Thomas Neumann, and Veit Köppen, editors, *Datenbanksysteme für Business, Technologie und Web (BTW), - Workshopband, 15. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), 11.-15.3.2013 in Magdeburg, Germany. Proceedings*, volume 216 of *LNI*, pages 181–195. GI, 2013. URL http://www.btw-2013.de/proceedings/Enhancing%20Named%20Entity%20Extraction%20by%20Effectively%20Incorporating%20the%20Crowd.pdf.

[9] Yu Cheng, Zhengzhang Chen, Jiang Wang, Ankit Agrawal, and Alok N. Choudhary. Bootstrapping active name disambiguation with crowdsourcing. In Qi He, Arun Iyengar, Wolfgang Nejdl, Jian Pei, and Rajeev Rastogi, editors, *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 1213–1216. ACM, 2013. DOI https://doi.org/10.1145/2505515.2507858.

[10] Nancy A. Chinchor. Overview of MUC-7/MET-2. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*. NIST, 1998. URL http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/overview.html.

[11] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1):37–46, 1960. DOI https://doi.org/10.1177/001316446002000104.

[12] Trevor Cohn and Lucia Specia. Modelling annotator bias with multi-task gaussian processes: An application to machine translation quality estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 32–42. Association for Computational Linguistics, 2013. URL http://aclweb.org/anthology/P/P13/P13-1004.pdf.

[13] Edward L. Deci and Richard M. Ryan. *Intrinsic Motivation and Self-Determination in Human Behavior*. Perspectives in Social Psychology. Springer, 1985. DOI https://doi.org/10.1007/978-1-4899-2271-7.

[14] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. ZenCrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In Alain Mille, Fabien L. Gandon, Jacques Misselis, Michael Rabinovich, and Steffen Staab, editors, *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*, pages 469–478. ACM, 2012. DOI https://doi.org/10.1145/2187836.2187900.

[15] Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. Analysis of named entity recognition and linking for tweets. *Information Processing and Management,*, 51(2):32–49, 2015. DOI https://doi.org/10.1016/j.ipm.2014.10.006.

[16] Leon Derczynski, Kalina Bontcheva, and Ian Roberts. Broad Twitter corpus: A diverse named entity recognition resource. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 1169–1179. ACL, 2016. URL http://aclweb.org/anthology/C/C16/C16-1111.pdf.

[17] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In Ricardo A. Baeza-Yates, Stefano Ceri, Piero Fraternali, and Fausto Giunchiglia, editors, *Proceedings of the First International Workshop on Crowdsourcing Web Search, Lyon, France, April 17, 2012*, volume 842 of *CEUR Workshop Proceedings*, pages 26–30. CEUR-WS.org, 2012. URL http://ceur-ws.org/Vol-842/

crowdsearch-difallah.pdf.

[18] George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie Strassel, and Ralph M. Weischedel. The automatic content extraction (ACE) program - tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*. European Language Resources Association, 2004. URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf.

[19] Oluwaseyi Feyisetan, Elena Simperl, Ramine Tinati, Markus Luczak-Rösch, and Nigel Shadbolt. Quick-and-clean extraction of linked data entities from microblogs. In Harald Sack, Agata Filipowska, Jens Lehmann, and Sebastian Hellmann, editors, *Proceedings of the 10th International Conference on Semantic Systems, SEMANTICS 2014, Leipzig, Germany, September 4-5, 2014*, pages 5–12. ACM, 2014. DOI https://doi.org/10.1145/2660517.2660527.

[20] Oluwaseyi Feyisetan, Elena Simperl, Max Van Kleek, and Nigel Shadbolt. Improving paid microtasks through gamification and adaptive furtherance incentives. In Aldo Gangemi, Stefano Leonardi, and Alessandro Panconesi, editors, *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pages 333–343. ACM, 2015. DOI https://doi.org/10.1145/2736277.2741639.

[21] Tim Finin, William Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark 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, Los Angeles, June 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W10-0713.

[22] Hege Fromreide, Dirk Hovy, and Anders Søgaard. Crowdsourcing and annotating NER for Twitter #drift. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 2544–2547. European Language Resources Association (ELRA), 2014. URL http://www.lrec-conf.org/proceedings/lrec2014/summaries/421.html.

[23] Alec Go, Lei Huang, and Richa Bhayani. Twitter sentiment analysis, 2009. URL https://nlp.stanford.edu/courses/cs224n/2009/fp/3.pdf. CS224N - Final Project Report.

[24] Annika Hinze, Ralf Heese, Markus Luczak-Rösch, and Adrian Paschke. Semantic enrichment by non-experts: Usability of manual annotation tools. In Philippe Cudré-Mauroux, Jeff Heflin, Evren Sirin, Tania Tudorache, Jérôme Euzenat, Manfred Hauswirth, Josiane Xavier Parreira, Jim Hendler, Guus Schreiber, Abraham Bernstein, and Eva Blomqvist, editors, *The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I*, volume 7649 of *Lecture Notes in Computer Science*, pages 165–181. Springer, 2012. DOI https://doi.org/10.1007/978-3-642-35176-1_11.

[25] Annika Hinze, Ralf Heese, Alexa Schlegel, and Markus Luczak-Rösch. User-defined semantic enrichment of full-text documents: Experiences and lessons learned. In Panayiotis Zaphiris, George Buchanan, Edie Rasmussen, and Fernando Loizides, editors, *Theory and Practice of Digital Libraries - Second International Conference, TPDL 2012, Paphos, Cyprus, September 23-27, 2012. Proceedings*, volume 7489 of *Lecture Notes in Computer Science*, pages 209–214. Springer, 2012. DOI https://doi.org/10.1007/978-3-642-33290-6_23.

[26] Nolan Lawson, Kevin Eustice, Mike Perkowitz, and Meliha Yetisgen-Yildiz. Annotating large email datasets for named entity recognition with mechanical turk. In Chris Callison-Burch and Mark Dredze, editors, *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 71–79, Los Angeles, June 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W10-0712.

[27] Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. Recognizing named entities in tweets. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 359–367. Association for Computational Linguistics, 2011. URL http://www.aclweb.org/anthology/P11-1037.

[28] Markus Luczak-Rösch and Ralf Heese. Linked data authoring for non-experts. In Christian Bizer, Tom Heath, Tim Berners-Lee, and Kingsley Idehen, editors, *Proceedings of the WWW2009 Workshop on Linked Data on the Web, LDOW 2009, Madrid, Spain, April 20, 2009.*, volume 538 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009. URL http://ceur-ws.org/Vol-538/ldow2009_paper4.pdf.

[29] Mónica Marrero, Sonia Sánchez-Cuadrado, Jorge Morato, and Yorgos Andreadakis. Evaluation of named entity extraction systems. In Alexander Gelbukh, editor, *Advances in Computational Linguistics*, volume 41 of *Research in Computing Science*, pages 47–58. Instituto Politécnico Nacional, 2009. URL http://www.cicling.org/2009/RCS-41/047-058.pdf.

[30] Diana Maynard, Kalina Bontcheva, and Hamish Cunningham. Towards a semantic extraction of named entities. In Ruslan Mitkov, editor, *Proceedings of RANLP - 2003 (Recent Advances in Natural Language Processing) 10-12 September 2003, Borovets, Bulgaria*. BAS, 2003.

[31] Einat Minkov, Richard C. Wang, and William W. Cohen. Extracting personal names from email: Applying named entity recognition to informal text. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*, pages 443–450. Association for Computational Linguistics, 2005. URL http://aclweb.org/anthology/H/H05/H05-1056.pdf.

[32] Robert Morris. Crowdsourcing workshop: The emergence of affective crowdsourcing. In Michael Bernstein, Björn Hartmann, Ed H. Chi, Niki Kittur, Lydia Chilton, and Robert C. Miller, editors, *CHI 2011 Workshop on Crowdsourcing and Human Computation: Systems, Studies, and Platforms, May 8, 2011*, 2011. URL http://www.humancomputation.com/crowdcamp/chi2011/papers/morris.pdf.

[33] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, 30

(1):3–26, 2007. DOI https://doi.org/10.1075/li.30.1.03nad.

[34] Stefanie Nowak and Stefan M. Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In James Ze Wang, Nozha Boujemaa, Nuria Oliver Ramirez, and Apostol Natsev, editors, *Proceedings of the 11th ACM SIGMM International Conference on Multimedia Information Retrieval, MIR 2010, Philadelphia, Pennsylvania, USA, March 29-31, 2010*, pages 557–566. ACM, 2010. DOI https://doi.org/10.1145/1743384.1743478.

[35] Barbara Plank, Dirk Hovy, and Anders Søgaard. Learning part-of-speech taggers with inter-annotator agreement loss. In Gosse Bouma and Yannick Parmentier, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 742–751. Association for Computational Linguistics, 2014. URL http://aclweb.org/anthology/E/E14/E14-1078.pdf.

[36] 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 *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1713–1722. Association for Computational Linguistics, 2013. URL http://aclweb.org/anthology/P/P13/P13-1168.pdf.

[37] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1524–1534. Association for Computational Linguistics, 2011. URL http://www.aclweb.org/anthology/D11-1141.

[38] Giuseppe Rizzo and Raphaël Troncy. NERD: evaluating named entity recognition tools in the web of data. In James Fan and Aditya Kalyanpur, editors, *Proceedings of Workshop on Web Scale Knowldge Extraction (WEKEX'11) co-located with the 10th International Semantic Web Conference (ISWC 2011), October 24, 2011, Bonn, Germany*, 2011. URL http://www.eurecom.fr/publication/3517.

[39] Marta Sabou, Arno Scharl, and Michael Föls. Crowdsourced knowledge acquisition: Towards hybrid-genre workflows. *International Journal on Semantic Web and Information Systems*, 9(3):14–41, 2013. DOI https://doi.org/10.4018/ijswis.2013070102.

[40] Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. Corpus annotation through crowdsourcing: Towards best practice guidelines. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 859–866. European Language Resources Association (ELRA), 2014. URL http://www.lrec-conf.org/proceedings/lrec2014/summaries/497.html.

[41] Hassan Saif, Yulan He, and Harith Alani. Semantic sentiment analysis of twitter. In Philippe Cudré-Mauroux, Jeff Heflin, Evren Sirin, Tania Tudorache, Jérôme Euzenat, Manfred Hauswirth, Josiane Xavier Parreira, Jim Hendler, Guus Schreiber, Abraham Bernstein, and Eva Blomqvist, editors, *The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I*, volume 7649 of *Lecture Notes in Computer Science*, pages 508–524. Springer, 2012. DOI https://doi.org/10.1007/978-3-642-35176-1_32.

[42] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147. Association for Computational Linguistics, 2003. URL http://aclweb.org/anthology/W/W03/W03-0419.pdf.

[43] Elena Simperl, Roberta Cuel, and Martin Stein. *Incentive-Centric Semantic Web Application Engineering*. Synthesis Lectures on The Semantic Web. Morgan & Claypool Publishers, 2013. DOI https://doi.org/10.2200/S00460ED1V01Y201212WBE004.

[44] Vivek Kumar Singh, Rajesh Piryani, Ashraf Uddin, and Pranav Waila. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In *Proceedings, 2013 IEEE International Multi Conference on Automation, Computing, Control, Communication and Compressed Sensing, 22 & 23rd Feb 2013, iMac4s - 2013*, pages 712–717. IEEE, 2013. DOI https://doi.org/10.1109/iMac4s.2013.6526500.

[45] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 254–263. Association for Computational Linguistics, 2008. URL http://www.aclweb.org/anthology/D08-1027.

[46] Beth Trushkowsky, Tim Kraska, Michael J. Franklin, and Purnamrita Sarkar. Crowdsourced enumeration queries. In Christian S. Jensen, Christopher M. Jermaine, and Xiaofang Zhou, editors, *29th IEEE International Conference on Data Engineering, ICDE 2013, Brisbane, Australia, April 8-12, 2013*, pages 673–684. IEEE Computer Society, 2013. DOI https://doi.org/10.1109/ICDE.2013.6544865.

[47] Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Michael Röder, Daniel Gerber, Sandro Athaide Coelho, Sören Auer, and Andreas Both. AGDISTIS - Graph-based disambiguation of named entities using linked data. In Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig A. Knoblock, Denny Vrandecic, Paul T. Groth, Natasha F. Noy, Krzysztof Janowicz, and Carole A. Goble, editors, *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, volume 8796 of *Lecture Notes in Computer Science*, pages 457–471. Springer, 2014. DOI https://doi.org/10.1007/978-3-319-11964-9_29.

[48] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. Community-based Bayesian aggregation models for crowdsourcing. In Chin-Wan Chung, An-

drei Z. Broder, Kyuseok Shim, and Torsten Suel, editors, *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014*, pages 155–164. ACM, 2014. DOI https://doi.org/10.1145/2566486.2567989.

[49] Luis von Ahn and Laura Dabbish. Designing games with a purpose. *Communications of the ACM (CACM)*, 51(8):58–67, 2008. DOI https://doi.org/10.1145/1378704.1378719.

[50] Robert Voyer, Valerie Nygaard, Will Fitzgerald, and Hannah Copperman. A hybrid model for annotating named entity training corpora. In Nianwen Xue and Massimo Poesio, editors, *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 243–246, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W10-1839.

[51] Aobo Wang, Vu Cong Duy Hoang, and Min-Yen Kan. Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*, 47(1):9–31, 2013. DOI https://doi.org/10.1007/s10579-012-9176-1.

[52] Meliha Yetisgen-Yildiz, Imre Solti, Fei Xia, and Scott Halgrim. Preliminary experiments with amazon's mechanical turk for annotating medical named entities. In Chris Callison-Burch and Mark Dredze, editors, *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 180–183, Los Angeles, June 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W10-0728.

[53] Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. A survey of crowdsourcing systems. In *PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, MA, USA, 9-11 Oct., 2011*, pages 766–773. IEEE, 2011. DOI https://doi.org/10.1109/PASSAT/SocialCom.2011.203.