Harnessing Diversity in Crowds and Machines for Better NER Performance

Oana Inel and Lora Aroyo

Vrije Universiteit Amsterdam, The Netherlands
oana.inel@vu.nl, lora.aroyo@vu.nl

Abstract. Over the last years, information extraction tools have gained a great popularity and brought significant performance improvement in extracting meaning from structured or unstructured data. For example, named entity recognition (NER) tools identify types such as people, organizations or places in text. However, despite their high F1 performance, NER tools are still prone to brittleness due to their highly specialized and constrained input and training data. Thus, each tool is able to extract only a subset of the named entities (NE) mentioned in a given text. In order to improve *NE Coverage*, we propose a hybrid approach, where we first aggregate the output of various NER tools and then validate and extend it through crowdsourcing. The results from our experiments show that this approach performs significantly better than the individual state-of-the-art tools (including existing tools that integrate individual outputs already). Furthermore, we show that the crowd is quite effective in *(1)* identifying mistakes, inconsistencies and ambiguities in currently used ground truth, as well as in *(2)* a promising approach to gather ground truth annotations for NER that capture a multitude of opinions.

Keywords: crowdsourcing, disagreement, diversity, perspectives, opinions, named entity extraction, named entity typing, hybrid machine-crowd workflow, crowdsourcing ground truth

1 Introduction

Named entity recognition (NER) is a powerful information extraction (IE) technique for identifying named entities (NEs) such as people, places, organizations, events and, to some extent, numerical values or time periods. Nowadays, there is an abundance of off-the-shelf NER tools [1]. When compared however, their output significantly varies in terms of: *(1)* the existence of an entity, *(2)* the entity surface form (*i.e.*, entity span) and the entity type, *(3)* the knowledge base used for disambiguation, or *(4)* the confidence scores given for an entity. This makes it difficult to choose the best NER tool as they all seem to have a partially good and partially not so good performance.

Even though some NER tools have reached human-like performance, they are still highly dependent on the input type and ground truth (gold standard) [2]. For example, a NER tool trained on particular input types or entity types performs

well only on similar data. Research [3] has shown that NER tools trained on English news articles achieve an accuracy of 85%-90% on this type of data, but perform very poor on short, ill-formed texts, such as microblogs. Similarly, the quality and the size of the ground truth could bias NER towards a particular annotation perspective. In [2], the authors show that many NER tools have very low performance when dealing with the diversity of miscellaneous entity types.

The mainstream approach of gathering ground truth for NER is still by means of experts, who typically follow over-specified annotation guidelines to increase the *inter-annotator agreement* between experts. Such guidelines are known to be prone to denying the intrinsic language ambiguity and its multitude of perspectives and interpretations [4]. Thus, ground truth datasets might not always be 'gold' or 'true' in terms of capturing the real text meaning and interpretation diversity. More recent work has been focusing on capturing the *inter-annotator disagreement* [5] to provide a new type of ground truth, where language ambiguity is considered. As crowdsourcing has proven to be a reliable method for IE in various domains, *e.g.*, news [6], tweets [7] and more specialized tasks such as entity typing [8], there is an increasing number of hybrid NER approaches that combine machine and crowd-based IE [9]. However, they all suffer from the same *'lack of understanding of ambiguity'* as the traditional NER tools.

This paper aims to answer the following research question: *can we leverage the machine and crowd diversity to improve NER performance?*. We propose a hybrid multi-machine-crowd approach where state-of-the-art NER tools are combined and their aggregated output is validated and improved through crowdsourcing. We perform the crowdsourcing experiments in the context of the CrowdTruth approach [10] and methodology [5] that aims at capturing the inherent language ambiguity by means of disagreement. Thus, we argue that:

- **H1**: Aggregating the output of NER tools by harnessing the inter-tool disagreement (Multi-NER) performs better than the individual NERs (Single-NER); we experiment with existing Wikipedia sentence-based ground truth datasets and show that disagreement among NER improves their performance; we also show that the crowd is effective in spotting NER mistakes;
- **H2**: NER performance is influenced by the rigidness of the ground truth;
- **H3**: Crowdsourced ground truth by harnessing inter-annotator disagreement produces diversity in annotations and thus, improves the aggregated output of NER tools; we show that the crowd can produce a better ground truth.

The main contributions of this paper, besides addressing the above mentioned results, are: *(1)* a hybrid workflow for NER that improves significantly current NER by means of disagreement-based aggregation and crowdsourcing; *(2)* a method for improving ground truth datasets through fostering disagreement among the machines and crowd; *(3)* a data and NER tool agnostic method to improve the NE coverage, *i.e.*, can be used with any type or any number of NER tools and can be applied on any number and type of entities; *(4)* a disagreement-aware approach that effectively mitigates the issues of NER tools.

The paper is structured as follows. Section 2 introduces the use case and the datasets, while Section 3 covers the state of the art. Section 4 contains the

comparative analysis of multiple NER tools and their aggregated output. Section 5 outlines the crowdsourcing experimental setup. Further, Section 6 presents the crowdsourcing results, while Section 7 discusses the results. Finally, Section 8 concludes and introduces the future work.

2 Use Case and Datasets

We performed named entity extraction with five state-of-the-art NER tools: NERD-ML[1], TextRazor[2], THD[3], DBpediaSpotlight[4], and SemiTags[5]. NERD-ML [11] is an extension of NERD [12], a NER tools unifier, that uses machine learning for improved results. We performed a comparative analysis of *(1)* their performance (output) and *(2)* their combined performance (output), on two ground truth (GT) evaluation datasets used during Task 1 of the Open Knowledge Extraction (OKE) semantic challenge at ESWC in 2015[6] ($OKE2015$) and 2016[7] ($OKE2016$) respectively. Table 1 presents the summary of the datasets: in total, there are 156 Wikipedia sentences with 1007 annotated named entities of types *place*, *person*, *organization* and *role*.

Table 1: Datasets Overview

	OKE2015			***OKE2016***		
	Sentences	**Named Entities**		**Sentences**	**Named Entities**	
	101	*Place*	120	55	*Place*	44
		Person	304		*Person*	105
		Organization	139		*Organization*	105
		Role	103		*Role*	86
Total	101	664[8]		55	340	

3 Related Work

3.1 Open Knowledge Extraction Systems

The systems proposed during the OKE challenges have been evaluated on datasets described in Section 2. The ADEL system [13] had the best performance in 2015, with an F1-score of 0.60, by implementing a hybrid 3-steps approach that combines an off-the-shelf NER model together with POS-tagging, a linking step

[1] `http://nerd.eurecom.fr`
[2] `https://www.textrazor.com`
[3] `http://ner.vse.cz/thd/`
[4] `http://dbpedia-spotlight.github.io/demo/`
[5] `http://nlp.vse.cz/SemiTags/`
[6] `https://github.com/anuzzolese/oke-challenge`
[7] `https://github.com/anuzzolese/oke-challenge-2016`
[8] The sum per type is not equal to 664 because 2 entities have 2 distinct types.

through DBpedia and Wikipedia, and a pruning step for removing the entities that are out of scope. A second system, FRED [14], had a micro F1-score of 0.34 and a macro F1-score of 0.22. However, the lower performance is due to the fact that the system was used with its default settings, without being adapted for this challenge. Similarly, the third participating system, FOX [15], is an off-the-self system. The system is not able to recognize the type *role*, thus, the F1-score is around 0.49. The enhanced version of ADEL [16] combines different models to improve the entity recognition and entity linking. The system described in [17] applies filtering and merging heuristics on the combined output of NER tools and semantic annotators. It outperforms ADEL with an F1-score above 0.65.

3.2 Crowdsourcing Named Entities

Crowdsourcing proved to be effective in gathering data semantics for various tasks, such as medical relation extraction [18], temporal events ordering [19,20], entity salience [21]. State-of-the-art NER tools have good performance when tested on news articles, but perform very poor on microblogs [3]. Thus, crowdsourcing has been used as an alternative to identify named entities in tweets [7,22]. When dealing with crowdsourced data, the quality plays an important role. Typical solutions for assessing the quality of crowdsourced data are based on the hypothesis [23] that there is only one right answer. However, we operate under the assumption that the disagreement among workers is not noise, but a signal [24,5] of *(i)* input ambiguity, *(ii)* worker quality and *(iii)* task clarity. Therefore, we run our crowdsoucing experiments on the CrowdTruth [10] framework.

3.3 Multi-NER, Hybrid Named Entity Recognition

Harnessing the agreement among NER tools proved to be effective in [25], since entities missed by one NER can be extracted by another NER. Agreement among NER tools is well captured by majority vote systems [26]. However, this could cut off relevant information such as, information supported by only one extractor and cases with more than one solution. When dealing with data on heterogeneous topics and domains, the accuracy of extracting named entities has been shown to increase when NER tools are combined [25,27].

In [9] the need of designing hybrid approaches for NER pipelines is stressed, based on the reliable crowd performance when identifying named entities in tweets. Systems that integrate machines and crowd have been already developed [6,28]. On the one hand, in [6], the authors propose a probabilistic model to choose the most relevant data that needs to be annotated by the crowd, in a hybrid machine-crowd approach. On the other hand, the crowdsourcing component has been integrated as a plugin in the GATE framework [28], but they still assume there is only one correct answer. Hybrid expert-crowd approaches [29] have also been envisioned. The authors optimize in time and cost the process of gathering expert annotations by involving the crowd: the experts mark the named entities, while the crowd provides the type of the entities.

4 Single-NER vs. Multi-NER Comparison

In this section we introduce the ***Multi-NER approach***, an approach that combines the output of five state-of-the-art NER tools. The NER tools whose output we combine are mentioned in Section 2. On the one hand, by performing a comparative analysis of the five individual NER tools and their combined output (Multi-NER), we aim to validate **H1**. On the other hand, by performing an empirical analysis of the cases where NER tools perform poorly, we aim to identify the factors that influence their performance (**H2**).

4.1 Single-NER vs. Multi-NER - entity surface

According to [2], the performance of each state-of-the-art NER differs on a dataset due to the fact that each NER tool uses different training data and different learning algorithms. However, evaluating the disagreement among them [25] proves to be effective in generating better outcomes. First, we compare the five Single-NERs and Multi-NER on the GT in Table 1, by looking at the entity surface. For this analysis we use all the NEs in the GT and all their alternatives, *i.e.*, all the surface forms for each entity in the GT, extracted by any NER tool. Considering this, we measure the following:

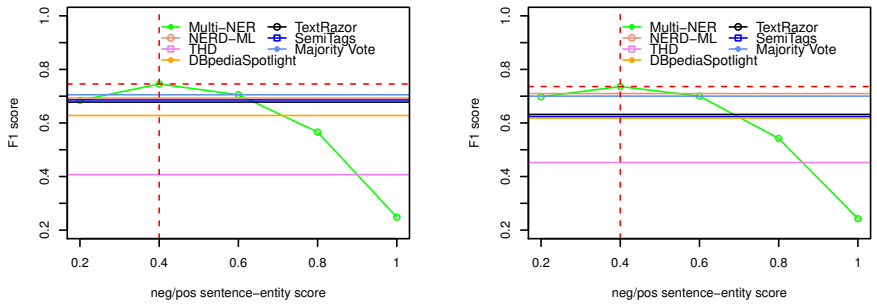
- *true positive (TP)*: the NE has the same surface form and the same offsets as the NE in the GT;
- *false positive (FP)*: the NE is only a partial overlap with the NE in the GT;
- *false negative (FN)*: the NEs in the GT that were not extracted by any NER, nor the Multi-NER.

Table 2: NER evaluation at the level of entity surface

	*OKE*2015						*OKE*2016					
	TP	FP	FN	P	R	F1	TP	FP	FN	P	R	F1
NERD-ML	401	93	263	0.812	0.604	**0.693**	209	37	131	0.85	0.615	**0.713**
SemiTags	366	**37**	298	**0.908**	0.551	0.686	161	**14**	179	**0.92**	0.474	0.625
THD	199	114	465	0.636	0.3	0.407	122	73	218	0.626	0.359	0.456
DBpediaSpotlight	411	234	253	0.637	0.619	0.628	**228**	119	**112**	0.657	**0.671**	0.664
TextRazor	**431**	177	**232**	0.709	**0.65**	0.678	207	105	133	0.663	0.609	0.635
Multi-NER	**555**	**403**	**109**	**0.579**	**0.836**	**0.684**	**299**	**218**	**41**	**0.578**	**0.879**	**0.698**

The comparison is shown in Table 2. Overall, there are high differences in terms of the number of TP, FP and FN cases for each state-of-the-art NER, but their performance in F1-score is still very similar. Although it seems that NERD-ML performs the best in F1-score across the two datasets, when looking at the exact numbers, we observe that the Multi-NER approach covers a significantly larger pool of entities, *i.e.*, has a significantly higher number of TPs and also a significantly lower number of FNs. However, the combined output of the NER tools also introduces a lot of FPs, but this only slightly decreases its performance.

The reason for the increased number of FPs is due to the high disagreement between the NER tools on the surface form of the entity (*i.e.*, the NER tools do



(a) OKE2015 (b) OKE2016

Fig. 1: Annotation quality F1 per negative/positive sentence-entity threshold

not agree on the exact entity span). On the one side, Multi-NER has a higher recall (with about 30%) on both datasets compared to TextRazor, the tool with the highest recall. This proves that the Multi-NER approach retrieves a higher number of relevant entities. On the other side, the low precision indicates the fact that many entities retrieved are not correct. Thus, our focus should be on improving the precision of the Multi-NER approach, while keeping a high recall.

To show that combining NER output and harnessing the diversity among them is beneficial, we applied the CrowdTruth methodology [10]. First, we introduce a core metric, the *sentence-entity score* which shows the likelihood of an entity to be in the GT based on how many NER tools extracted it. The *sentence-entity score* is equal to the ratio of NER tools that extracted the entity. In Figure 1a and Figure 1b we plotted the F1-score values for each NER and the F1-score of the Multi-NER approach for each sentence-entity score threshold. We use the sentence-entity score as a threshold for differentiating between a positive and a negative named entity. At the threshold of 0.4, Multi-NER outperforms the rest of the tools. Using McNemar's test, the results show that the difference in performance between NERD-ML and Multi-NER at its best performing threshold is statistically significant ($OKE2015$: $p < 2.2e^{-16}$, $OKE2016$: $p < 3.247e^{-11}$).

We have also plotted the F1-score for the majority vote approach, a mainstream approach when combining multiple NER tools. In our case, the majority vote includes all the entities that were extracted by at least 3 NER (sentence-entity score $>= 0.6$). The difference of performance is also statistically significant for majority vote vs. Multi-NER ($OKE2015$: $p < 2.2e^{-16}$, $OKE2016$: $p < 7.025e^{-12}$). Overall, the Multi-NER outperforms the state-of-the-art NER tools at a sentence-entity score $>= 0.4$, which fosters the idea that disagreement is beneficial, and it also outperforms the majority vote approach.

4.2 Single-NER vs. Multi-NER - entity surface & entity type

To better understand the results of our Multi-NER approach, we focus on analyzing the cases where the NER tools underperform. Table 3a and Table 3b contain the combined NER evaluation at the entity surface based on the entity type. We show how the TP, FP and FN cases from Table 2 are distributed across the types of interest: *person*, *place*, *organization* and *role*. The remaining of the section focuses on analyzing the FN and FP cases.

Table 3: NER evaluation at the level of entity surface and entity type

(a) $OKE2015$

	TP					FP					FN				
	Place	People	Org	Role	*Total*	Place	People	Org	Role	*Total*	Place	People	Org	Role	*Total*
NERD-ML	90	142	106	65	403	22	21	42	17	102	30	162	33	38	263
SemiTags	100	168	100	0	368	16	2	19	2	39	20	136	39	103	298
THD	62	35	55	49	201	17	17	62	29	125	58	269	84	54	465
DBpedia-Spotlight	99	156	81	77	413	26	62	124	26	238	21	148	58	26	253
TextRazor	110	174	109	40	433	31	14	118	24	187	9	130	30	63	232
Multi-NER	116	219	130	92	558	**54**	**91**	**214**	**66**	425	**4**	**85**	**9**	**11**	108

(b) $OKE2016$

	TP					FP					FN				
	Place	People	Org	Role	*Total*	Place	People	Org	Role	*Total*	Place	People	Org	Role	*Total*
NERD-ML	40	47	71	51	209	1	3	30	6	40	4	58	34	35	131
SemiTags	36	57	67	1	161	5	2	7	1	15	8	48	38	85	179
THD	36	12	33	41	122	3	1	55	14	73	8	93	72	45	218
DBpedia-Spotlight	38	70	56	64	228	5	7	93	14	119	6	35	49	22	112
TextRazor	36	57	83	31	207	15	4	79	12	110	8	48	22	55	133
Multi-NER	44	78	100	77	299	**21**	**13**	**157**	**34**	225	**0**	**27**	**5**	**9**	41

4.3 Analysis of False Negative Named Entities

We started with a manual inspection of the FN cases in order to understand which are the NEs that the NER tools fail to identify. Typically, by using the Multi-NER approach, it is natural to have high recall values and lower precision values. However, in both Table 3a and Table 3b we see that there are many entities of type *person* that are missed ($OKE2015$ recall - 0.72 and $OKE2016$ recall - 0.74). When analyzing in detail, we identify three main problems:

- the NER tools have problems in identifying coreferences, or identifying personal and possessive pronouns as named entities
  - o in $DS2015$, there were 26/27 such cases
  - o in $DS2016$, there were 83/85 such cases
- there are errors in the ground truth: in $OKE2015$, "One of the them", which is a clear mistake, is considered a correct named entity

- ambiguous combination of type *role* and *people*, *e.g.*, "Bishop Petronius", "Queen Elizabeth II"; "Bishop Petronius" is a *person*, while "Queen Elizabeth II" is a false entity, because "Queen" - *role* and "Elizabeth II" - *people*.

The type *place* seems to be the one that has the lowest number of FNs: in $OKE2016$ all the entities of type *place* were extracted, while in $OKE2015$ only four cases were missed. Here, we identify one main issue, in all four cases the entity is a concatenation of multiple entities of type *place*. Furthermore, in $2/4$ cases the ground truth contains errors - the extracted entity span does not match with the given offsets As a general rule, in the $OKE2015$ dataset, the cases "City, Country" were extracted as a single entity of type *place*. However, the annotation guidelines for $OKE2016$ seemed to have changed, since all such cases were considered two different entities of type *place*. Thus, we observe that there is disagreement across the two ground truth datasets.

For the types *organization* and *role*, the general observation is that there is a high disagreement between the single NER tools and they constantly seem to have a high rate of FN for such entities. However, when looking at the Multi-NER approach, we see that overall, only a few cases were missed which means that at least one NER was able to extract the correct entity span. When looking in depth at the entities of type *organization* that were missed, we see two cases:

- in $OKE2016$ the entities missed were actually common entities in $5/5$ cases (*e.g.*, "state", "university", "company");
- in $OKE2015$ the entities missed were not common entities, but the GT:
  - o contains errors in $2/9$ cases (*e.g.*, "Sheffiel", "The Imperial Cancer Research Fund")
  - o contains non-English named-entities in $1/9$ cases
  - o contained combinations of named and common entities in $4/9$ cases (*e.g.*, "Boston Brahmin family", "Geiger's staff")

Since the entities of type *role* are common entities, the main issue of the NER tools is the fact that they extract other span alternatives instead of the one in the ground truth. Furthermore, in $OKE2015$ we had a French entity, while in $OKE2016$, in $5/9$ cases the entities were highly ambiguous, such as "membership", "originators". Looking further in the FN cases, we see that there are many ambiguities. For example, in $OKE2015$, we have the word phrase "Italian Jewish", where "Italian" is classified as a *person* and "Jewish" as a *role*. In another example for the same dataset, the word phrase "Hungarian Jews" is classified alone as an *organization*, while, in $OKE2016$ we find the word phrase "Jewish mother", where "Jewish" has no type and "mother" is a *role*. We see such inconsistencies across types as well: "independent school" is an incorrect *organization* type, but "independent contractor" is a correct entity of type *role*.

4.4 Analysis of False Positive Named Entities

We performed a similar manual evaluation on the FPs in order to understand how we can correct the results of the NER tools and improve the precision of the Multi-NER approach. For both datasets the precision of extraction an entity of type *organization* is significantly low ($OKE2015$ - 0.37, $OKE2016$ - 0.38). This

is due to the large number of FP cases, or in other words, the various alternatives for a single entity. The large majority of entities of type *organization* are combinations of *organization* and *place* (*e.g.*, "University of Rome") or combinations of *people* and *organization* (*e.g.*, "Niels Bohr Institute"). Thus, for each such entity, there are at least 2 more FP alternatives that are extracted.

The next type with many FPs is the type *person*. Here we identify:

- the NER tools usually extract correctly the name of the person, but they also extract partial matches (in $OKE2015$ we have 86/91 such partial matches, while in $OKE2016$ we have 11/13 such cases); when checking these cases we observe that the names that contain abbreviations, *e.g.*, "J. Hans D. Jensen", are the most prone to get any possible combination of the names;
- the NER tools extract combinations of *role* and *person*, especially when the *person* is an ethnic group (*e.g.*, "French author", "Canadian citizen");
- in $OKE2015$ we also find a combination of *place* and *person*, due to the ambiguity of the sentence (*e.g.*, "Turin Rita Levi-Montalcini"), which lacks a comma after the word "Turin".

Similarly, for the FP cases on type *place*, in the majority of the cases we identify partial overlaps with the entity in the ground truth, concatenated or nested locations. Moreover, we find:

- combinations of entities of type *role* and entities of type *place* (1/54 cases in $OKE2015$ and 3/21 cases in $OKE2016$)
- combinations of entities of type *organization* and entities of type *place* (5/54 cases in $OKE2015$ and none in $OKE2016$)

The type *role* is the most ambiguous, especially because these entities are common entities. The main issue of the NER tools is the precision of extracting such entities. Usually, they tend to extract both the most general word phrase that refers to a *role* (*e.g.*, "professor" instead of "assistant professor"), but also the most specific one (*e.g.*, "first black president" instead of "president").

5 Experimental Setup

The aim of our crowdsourcing experiments is two-fold. On the one hand we want to prove that the crowd is able to correct the mistakes of the NER tools. On the other hand, we want to show that the crowd can identify the ambiguities in the GT, which leads to a better NER pipeline performance and an improved GT.

5.1 Crowdsourcing Experimental Data

Our goal is to decrease both the number of false positive and false negative NEs through gathering a crowd-driven ground truth. To achieve this, we select every entity in the ground truth for which the NER tools provided alternatives. It is important to mention that we do not focus on identifying new entities, but only on correcting the ones that exist. Thus, we have the following two cases:

- *Crowd reduces the number of FP*: For each named entity in the ground truth that has multiple span alternatives we create an entity cluster. We also

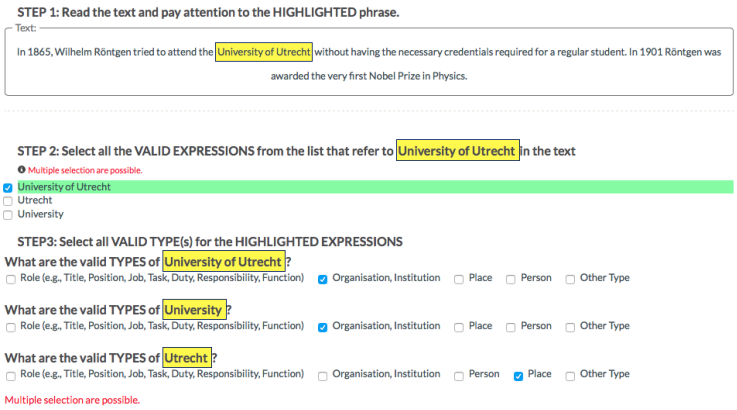


Fig. 2: Crowdsourcing Annotation Task

add the largest span among all the alternatives. For example, 'University of Rome' cluster is composed of: 'University', 'Rome' and 'University of Rome', where all entities have been extracted by at least one NER tool.

- *Crowd reduces the number of FN*: For each named entity in the GT that was not extracted, we create an entity cluster that contains the FN named entity and the alternatives returned by the NERs. Further, we add every other combination of words contained in all the alternatives. This step is necessary because we do not want to introduce bias in the task, *i.e.*, the crowd should see all the possibilities, not only the expected one. For example, the entity 'fellow students' was not extracted by any of the NER tools. Instead, they extracted 'fellow' and 'students'. The entity cluster in this case is composed of: 'fellow students', 'fellow' and 'students'.

5.2 Crowdsourcing Annotation Task

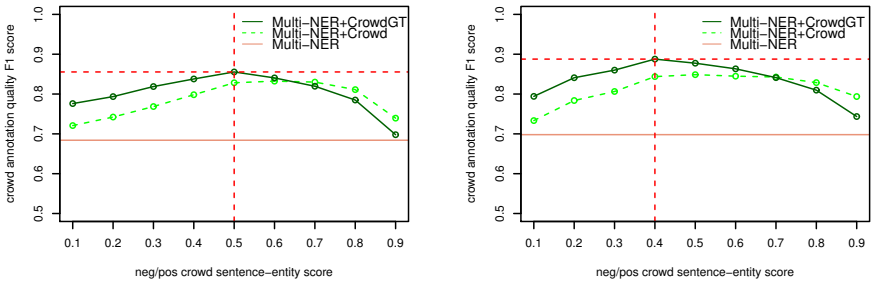
For both cases introduced in Section 5.1 we designed the same crowdsourcing task on CrowdFlower[9]. The overview of the task is presented in Figure 2. The goal of the crowdsourcing task is two-fold: *(i)* identification of valid expressions from a list that refer to a highlighted phrase in yellow (Step 2 in Figure 2 and *(ii)* selection of the type for each expression in the list, from a predefined set of choices - *place*, *person*, *organization*, *role* and *other* (Step 3 in Figure 2).

The input for this crowdsourcing task consists of *(i)* a sentence from either $OKE2015$ or $OKE2016$, and *(ii)* a list of expressions that could potentially refer to a named entity. The list of expressions was created using the rules described in Section 5.1. In total, we ran 303 such pairs, distributed in 7 crowdsourcing jobs. The settings and the distribution per dataset is shown in Table 4.

[9] `www.crowdflower.com`

Table 4: Experimental Setup for Crowdsourcing Annotations

	Units	Jobs	Judg/Unit	Max Judg/Worker	Worker Country	Worker Level	Units/Page	Pay/Unit
OKE2015	202	2	15	15	UK, USA, AUS, CAN	3	1	2
OKE2016	101	5						



(a) OKE2015 (b) OKE2016

Fig. 3: Annotation quality F1 per neg/pos crowd sentence-entity threshold

5.3 CrowdTruth Metrics

We evaluate the crowdsourced data using the CrowdTruth methodology and metrics [10], by adapting the core CrowdTruth metric, the sentence-relation score [24]. In our case, we measure *(1) crowd sentence-entity score* - the likelihood of a sentence to contain a valid entity expression and *(2) crowd entity-type score* - the likelihood of an expression to refer to the given types. These scores are computed using the cosine similarity measure. To identity the low-quality workers we apply two CrowdTruth worker metrics [10], the worker-worker agreement and the worker cosine. These measures indicate how much a worker disagrees with the rest of the workers on the units they solved in common and across the entire dataset. Low values for both metrics mean that the workers consistently disagree with the rest of the workers. Their annotations are thus removed.

6 Results

This section presents the crowdsourcing results[10], with focus on analyzing the added value of using the crowd in hybrid Multi-NER pipelines. In short, we gathered 4,545 judgments, from a total of 464 workers. After applying the CrowdTruth metrics, we identified 108 spammers, that contributed to a total of 1,172 low-quality annotations, which were removed from the final data.

We plotted in Figures 3a and 3b the F1-score values at each crowd sentence-entity score, as described in Section 5.3. When compared with the ground truth,

[10] `http://data.crowdtruth.org/crowdsourcing-ne-goldstandards/`

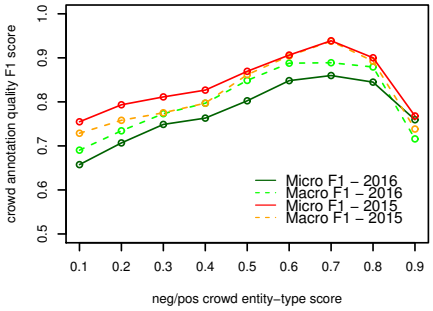


Fig. 4: Annotation quality F1 per neg/pos crowd entity-type threshold

we see that for each crowd sentence-entity score the crowd enhanced Multi-NER (Multi-NER+Crowd) performs much better than the Multi-NER approach. On the $OKE2015$ dataset the crowd performs the best at the crowd entity-score threshold of 0.6 with a F1-score of 0.832, while on $OKE2016$ the crowd has the best performance at a threshold of 0.5, with an F1-score of 0.848. This means that the crowd can correctly reduce the number of FPs. The difference is also statistically significant for both datasets. Using McNemar's test we get a p-value equal to $6.999e^{-07}$ for $OKE2015$ and p-value 0.01234 for $OKE2016$.

From these graphs, it is natural to assume that the crowd diversity in opinion is indeed not an indication of noise, but signal. In the analysis performed in Section 4 we observed that many entities in the ground truth are ambiguous and could have multiple interpretations. Thus, we performed a manual evaluation of the entities in the ground truth and allowed for a richer diversity. When the entities were ambiguous, *"professor" vs. "assistant professor", "Bishop Petronius" vs. "Bishop" vs. "Petronius"*, we included all the possible alternatives. In Figure 3 this evaluation is indicated by *Multi-NER+CrowdGT*, which stands for enhanced Multi-NER through crowd-driven ground truth gathering. Here we observe that we get even a higher performance ($OKE2015$ - F1 of 0.85 and $OKE2016$ - F1 of 0.88). For both datasets, we see that in this case the best performance threshold is consistently a fraction lower than the one when the crowd is compared with the experts.

We also evaluated the performance of the crowd on the entity types. For this evaluation we considered only the entities in the ground truth that have been used in the crowdsourcing tasks (227 entities in $OKE2015$ and 109 entities in $OKE2016$). Because we deal with multiple classes, in Figure 4 we plotted the macro F1-score and the micro F1-score based on the crowd weights, *i.e.*, based on the crowd entity-type score. Overall, the crowd is able to capture the correct entity type, as at each threshold all the F1 scores are higher than 0.65, with a maximum performance 0.93 for $OKE2015$ and 0.85 for $OKE2016$.

7 Discussion

Our first hypothesis was that a *Multi-NER approach performs better than a single NER.* As expected, when combining the output of multiple NER tools we increase the number of TP and decrease the number of FN. This observation is in agreement with the fact that in general, single NERs are trained with different data and through different approaches, *i.e.,* entities missed by one NER can be returned by another NER. Furthermore, the conventional belief when dealing with diversity is that *the more instances we have in agreement, the better.* To address this issue and prove the contrary, we follow and apply the CrowdTruth approach, *i.e.,* disagreement is not noise, but signal. In Figures 1a and 1b we can see that taking only the entities that have been extracted by many NER tools achieves a lower performance than most of the single NER. In contrast, the more disagreement we allow, the better our Multi-NER performs, which shows that a Multi-NER approach, overall, performs better than any single NER. Interestingly, although NERD-ML seems to overall outperform our approach, when leveraging the NERs diversity, at a 0.4 sentence-entity score threshold, we observe a statistically significant improvement for our method on both datasets.

The NER performance is influenced by the rigidity and the ambiguity of the GT, which can be proved by looking at the FN and FP cases. First, the annotation guidelines of the GT, do not seem to align with the GT used by the NER tools: *(1)* personal and especially possessive pronouns are not considered named entities by NER, in contrast to our GT; *(2)* the GT is inconsistent for the same dataset and across datasets; *(3)* the GT contains ambiguities that are fostered for difficult types such as *role*; *(4)* the GT contains errors. The NER tools tend to extract multiple span alternatives for an entity, while the GT does not allow for multiple perspectives on the entity span. We observe this cuts off meaningful data. Furthermore, many challenge submissions (Section 3.1) were off-the-shelf tools, GT-agnostic. The tools performance did not exceed an F1-score of 0.65, which is quite low given that we deal with well-formed English Wikipedia sentences. We argue that the GT ambiguity also impacts their performance.

Overall, the crowd improved the performance of the NER tools. In Figure 3 it is interesting to see that the best performing threshold for Multi-NER+Crowd is not only a pick, but it is an interval of thresholds (in $OKE2015$ - between 0.5 and 0.7, while in $OKE2016$ - between 0.4 and 0.7). Furthermore, we see that the lower end of the intervals is correlated with the best performing threshold for the crowd-driven ground truth (Multi-NER+CrowdGT). We believe this is an indicator of the fact that entities in that interval are more prone to be ambiguous. Thus, allowing for diversity provides better ground truth. For the type analysis it is interesting to see that the micro F1 and macro F1 differ on each dataset. This behavior is due to the highly unbalanced number of entities in each class for $OKE2016$, where we have 62 entities of type *organization* and only 7 of type *person*. Since in the case of micro averaging larger classes dominate smaller classes, for $OKE2016$ we should consider the macro F1 score as a better indicator. However, for $OKE2015$ the classes are more balanced, so we can give them the same weight, thus, the micro F1 is a better indicator.

8 Conclusion

In this paper we proposed a hybrid Multi-NER crowd-driven approach for improved NER performance. Following the CrowdTruth methodology - *disagreement is not noise but signal*, we showed the added value of leveraging the machines and crowd diversity in a 3-step approach. First, our Multi-NER approach, by considering the data ambiguity, has a significantly higher coverage of entities than Single-NER tools when compared to given ground truth. Furthermore, when leveraging the NERs diversity, we show a significant improvement over state-of-the-art Single-NER on both datasets. Second, through data inspection of the ground truth and the factors that answer for the increased number of false positive and false negative entities, we observed that the NER performance is highly dependent on the ambiguity and inconsistency of such ground truth datasets. Third, our evaluation has shown that the crowd, *by harnessing the inter-annotator disagreement*, is able to correct the mistakes of the NER tools by reducing the total number of false positive cases. Furthermore, the crowd-driven ground truth gathering, that harnesses diversity, perspectives and granularities, proves to be a more reliable way of creating a ground truth when dealing with the natural language ambiguity and the overall task ambiguity.

Although the current performance of the hybrid Multi-NER crowd-driven approach reaches high values, in future work we can focus on reducing the false negative cases related to personal and possessive pronouns. Furthermore, we can optimize the crowdsourcing approach in terms of time and cost by only validating and correcting the named entities with low confidence of being correct.

References

1. Gangemi, A.: A comparison of knowledge extraction tools for the semantic web. In: Extended Semantic Web Conference, Springer (2013) 351–366
2. Rizzo, G., van Erp, M., Troncy, R.: Benchmarking the extraction and disambiguation of named entities on the semantic web. In: LREC. (2014) 4593–4600
3. Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., Bontcheva, K.: Analysis of named entity recognition and linking for tweets. Information Processing & Management **51**(2) (2015) 32–49
4. Bayerl, P.S., Paul, K.I.: What determines inter-coder agreement in manual annotations? a meta-analytic investigation. Computational Linguistics **37**(4) (2011)
5. Aroyo, L., Welty, C.: Truth Is a Lie: CrowdTruth and 7 Myths about Human Computation. AI Magazine **36**(1) (2015)
6. Demartini, G., Difallah, D.E., Cudré-Mauroux, P.: Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In: Proceedings of the 21st international conference on WWW, ACM (2012) 469–478
7. Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., Dredze, M.: Annotating named entities in twitter data with crowdsourcing. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, ACL (2010) 80–88
8. Bu, Q., Simperl, E., Zerr, S., Li, Y.: Using microtasks to crowdsource DBpedia entity classification: A study in workflow design. Semantic Web Journal (2016)

9. Feyisetan, O., Luczak-Roesch, M., Simperl, E., Tinati, R., Shadbolt, N.: Towards hybrid NER: a study of content and crowdsourcing-related performance factors. In: European Semantic Web Conference, Springer (2015) 525–540
10. Inel, O., Khamkham, K., Cristea, T., Dumitrache, A., Rutjes, A., van der Ploeg, J., Romaszko, L., Aroyo, L., Sips, R.J.: CrowdTruth: machine-human computation framework for harnessing disagreement in gathering annotated data. In: The Semantic Web–ISWC 2014. Springer (2014) 486–504
11. Van Erp, M., Rizzo, G., Troncy, R.: Learning with the web: Spotting named entities on the intersection of nerd and machine learning. In: # MSM. (2013) 27–30
12. Rizzo, G., Troncy, R.: Nerd: a framework for unifying named entity recognition and disambiguation extraction tools. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the ACL, ACL (2012) 73–76
13. Plu, J., Rizzo, G., Troncy, R.: A hybrid approach for entity recognition and linking. In: Semantic Web Evaluation Challenge. (2015) 28–39
14. Consoli, S., Recupero, D.R.: Using FRED for named entity resolution, linking and typing for knowledge base population. In: Semantic Web Evaluation Challenge. (2015) 40–50
15. Röder, M., Usbeck, R., Speck, R., Ngomo, A.C.N.: Cetus–a baseline approach to type extraction. In: Semantic Web Evaluation Challenge. (2015) 16–27
16. Plu, J., Rizzo, G., Troncy, R.: Enhancing entity linking by combining NER models. In: Semantic Web Evaluation Challenge. (2016) 17–32
17. Chabchoub, M., Gagnon, M., Zouaq, A.: Collective disambiguation and semantic annotation for entity linking and typing. In: Semantic Web Evaluation Challenge. (2016) 33–47
18. Dumitrache, A., Aroyo, L., Welty, C.: Achieving expert-level annotation quality with CrowdTruth. (2015)
19. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In: Proceedings of EMNLP, Association for Computational Linguistics (2008) 254–263
20. Caselli, T., Sprugnoli, R., Inel, O.: Temporal information annotation: Crowd vs. experts. In: LREC (2016)
21. Inel, O., Caselli, T., Aroyo, L.: Crowdsourcing salient information from news and tweets. In: LREC. (2016) 3959–3966
22. Fromreide, H., Hovy, D., Søgaard, A.: Crowdsourcing and annotating ner for twitter# drift. In: LREC. (2014) 2544–2547
23. Nowak, S., Rüger, S.: How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In: Proceedings of the international conference on Multimedia information retrieval, ACM (2010)
24. Aroyo, L., Welty, C.: The Three Sides of CrowdTruth. Journal of Human Computation **1** (2014) 31–34
25. Chen, L., Ortona, S., Orsi, G., Benedikt, M.: Aggregating semantic annotators. Proceedings of the VLDB Endowment **6**(13) (2013) 1486–1497
26. Kozareva, Z., Ferrández, Ó., Montoyo, A., Muñoz, R., Suárez, A., Gómez, J.: Combining data-driven systems for improving named entity recognition. Data & Knowledge Engineering **61**(3) (2007) 449–466
27. Hellmann, S., Lehmann, J., Auer, S., Brümmer, M.: Integrating NLP using linked data. In: The Semantic Web–ISWC 2013. (2013) 98–113
28. Sabou, M., Bontcheva, K., Derczynski, L., Scharl, A.: Corpus annotation through crowdsourcing: Towards best practice guidelines. In: LREC. (2014) 859–866
29. Voyer, R., Nygaard, V., Fitzgerald, W., Copperman, H.: A hybrid model for annotating named entity training corpora. In: Proceedings of LAW IV, ACL (2010)