Improving NER with Diversity in Machines and Crowds

OANA INEL, Vrije Universiteit Amsterdam
LORA AROYO, Vrije Universiteit Amsterdam

1. INTRODUCTION

Named entity recognition (NER) is a powerful information extraction technique for identifying named entities (NEs) such as people, places, organizations, events and to some extent time periods. Even though there is an abundance of off-the-shelf NER tools [Gangemi 2013], each tool is able to extract only a subset of the NEs mentioned in a given text. NER tools would often 'disagree' on *(1)* the existence of an entity, *(2)* the entity surface and the entity type, *(3)* the knowledge base used for disambiguation, or *(4)* the confidence score given for an entity.

The mainstream approach of gathering ground truth for NER is still by means of experts. However, the guidelines followed by them are typically over-specified to increase the *inter-annotator agreement*. As a consequence, the intrinsic ambiguity of the language and its multitude of perspectives are denied [Bayerl and Paul 2011]. 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* [Aroyo and Welty 2015; Aroyo and Welty 2014] to provide a new type of ground truth, where language ambiguity is taken in consideration.

As crowdsourcing has proven to be a reliable method for IE in various domains, *e.g.*, news [Demartini et al. 2012], entity typing [Bu et al. 2016], there is an increasing number of hybrid NER approaches that combine machine and crowd-based IE [Feyisetan et al. 2015]. [Demartini et al. 2012] proposes a probabilistic model to choose the most relevant data to be annotated by the crowd, in a hybrid machine-crowd approach. The crowdsourcing component has been also integrated in the GATE framework [Sabou et al. 2014], or in hybrid expert-crowd workflows [Voyer et al. 2010]. However, they all suffer from the same *lack of understanding of ambiguity* as the traditional NER.

This paper aims to answer the following research question: *can we leverage machine and crowd diversity to improve NER performance by considering data ambiguity?*. We propose a hybrid multi-machine-crowd approach where state-of-the-art NER tools are combined (Multi-NER) and their aggregated output is validated and improved through crowdsourcing (Multi-NER + Crowd). We follow and apply the *CrowdTruth* approach, metrics and methodology [Inel et al. 2014; Aroyo and Welty 2014; Aroyo and Welty 2015] to overcome the usual limitations of the expert-based ground truth (GT).Thus, we propose: *(1)* a hybrid NER that improves significantly current NER by means of disagreement-based aggregation and crowdsourcing; *(2)* a method to improve the GT through fostering disagreement. Our experiments showed that *allowing for diversity provides better ground truth* and that *the crowd improved the performance of the NER tools.*

2. EXPERIMENTAL SETUP

We performed our experiments on the OKE Challenge 2016 dataset of 156 expert annotated English Wikipedia sentences with 1006 named entities with the following four types: *person* - 409, *place* - 164, *organization* - 244, *role* - 189. The data can be accessed at the following locations[1], under the name "testing dataset". This dataset also stands as expert ground truth (GT) in our analysis.

[1]https://github.com/anuzzolese/oke-challenge, https://github.com/anuzzolese/oke-challenge-2016

2.1 Single-NER vs. Multi-NER Performance

We started our experiments by extracting all the NEs from our dataset using five state-of-the-art NER tools (see first column of Table I). Next, we performed a comparative analysis of *(1)* the Single-NER tools performance and *(2)* their combined performance - Multi-NER, against our GT. In Table I we compared the performance in precision, recall and F1-score of the five state-of-the-art NER tools and our Multi-NER approach. The NERD-ML tool seems to perform the best in F1-score. However, Multi-NER has a significantly higher number of true positive (TP) entities (*i.e.*, the same span and offsets as in the GT) and a significantly lower number of false negative (FN) entities (*i.e.*, missed entities). The small decrease in F1-score for the Multi-NER approach is due to the significantly higher number of false positive (FP) entities (i.e., entities in the GT for which the NER tools returned only partial matches), due to the high disagreement between the NER tools tools on the surface form of the entity.

To show that combining NER output and harnessing diversity is beneficial, we applied the CrowdTruth methodology [Aroyo and Welty 2014]. 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, *i.e.*, the ratio of NER tools that extracted the entity. In Figure 1 we plotted the F1 values for each NER and the F1 of the Multi-NER approach at each sentence-entity score threshold. We use the sentence-entity score as a threshold for differentiating between a true and a false NE. At a threshold of 0.4, Multi-NER outperforms the NER tools and it is statistically significant, $p < 2.2e^{-16}$.

Table I. : Single-NER[2] vs. Multi-NER performance evaluation on entity surface

	TP	FP	FN	Precision	Recall	F1-score
NERD-ML	610	131	394	0.823	0.608	0.699
SemiTags	527	51	477	0.912	0.525	0.666
THD	321	187	683	0.632	0.320	0.425
DBpediaSpotlight	639	353	365	0.644	0.636	0.640
TextRazor	638	282	365	0.693	0.636	0.664
Multi-NER	854	621	150	0.579	0.851	0.689

In Table II we also evaluated the type of the entities in the GT. A manual inspection of the FP and FN cases identified some main problems with the **NER tools**: *(1)* do not identify coreferences, personal and possessive pronouns as NEs; *(2)* miss common entities, such as entities of type *role*; *(3)* usually extract correctly names of *people* and *organizations*, but they also extract all the possible partial matches of them, such as "University" and "Rome" in "University of Rome". Furthermore, the **ground truth**: *(1)* contains errors, misspelled words, *e.g.*, the NE "One of the them"; *(2)* contains ambiguous combinations of type *role* and *people*, *e.g.*, "Bishop Petronius" - *people*, while "Queen Elizabeth II" is not an entity, because "Queen" - *role* and "Elizabeth II" - *people*; *(3)* is not consistent, concatenations of multiple entities of type *place* are either considered a single entity or multiple.

2.2 Crowdsourcing for Better NER Performance

The aim of our crowdsourcing experiments is two-fold. On the one hand, we want to show that the *crowd is able to correct the mistakes of the NER tools*. On the other hand, we want to show that the *crowd is able to identify the ambiguities in the ground truth*, which leads to a better NER pipeline performance and an improved ground truth. Thus, we have the following two cases:

- *Crowd reduces the number of FP*: For each NE in the GT that has multiple span alternatives we create an entity cluster. We also add the largest span among all the alternatives.
- *Crowd reduces the number of FN*: For each NE in the GT that was not extracted, we create an entity cluster that contains the FN named entity and the alternatives returned by the NER. Further, we

[2]http://nerd.eurecom.fr,https://www.textrazor.com,http://ner.vse.cz/thd,http://dbpedia-spotlight.github.io,http://nlp.vse.cz/SemiTags

add every other combination of words contained in all the alternatives because the crowd should see all the possibilities, not only the correct one.

For both cases we designed the same crowdsourcing task on CrowdFlower[3]. The goal of the crowdsourcing task is two-fold: *(i)* identification of valid expressions from a list that refer to a highlighted phrase and *(ii)* selection of the type for each expression in the list, from a predefined set of choices - *place*, *people*, *organization*, *role* and *other*. The input for this crowdsourcing task consists of a sentence and a list of expressions that could potentially refer to a NE. In total, we ran 303 such pairs, distributed in 7 crowdsourcing jobs. We asked 15 judgments for each unit and each judgment is paid with 2 cents. We used level 3 workers (cf. CrowdFlower) from English speaking countries.

3. RESULTS AND CONCLUSION

This section presents the crowdsourcing results[4], with focus on analyzing the added value of using the crowd in hybrid Multi-NER pipelines. We gathered 4,545 judgments, from a total of 464 workers. We evaluated the crowdsourced data using the CrowdTruth methodology and metrics [Inel et al. 2014], by adapting the core CrowdTruth metric, the sentence-relation score [Aroyo and Welty 2014; Inel et al. 2014]. In our case, we measure the *crowd sentence-entity score* as the likelihood of a sentence to contain a valid entity expression. This scores are computed using the cosine similarity measure. After applying the CrowdTruth worker metrics, *i.e.*, worker cosine and worker-worker agreement, we identified 108 spammers, that contributed to a total of 1,172 low-quality annotations, which were then removed.

Figure 2 shows the F1 values at each crowd sentence-entity score threshold. We see that for each crowd sentence-entity score threshold the crowd enhanced Multi-NER (Multi-NER+Crowd) performs much better than Multi-NER. The crowd performs the best at the crowd entity-score threshold of 0.6 with a F1 of 0.836, and thus, the crowd can correctly eliminate FPs. The difference is also statistically significant, $p = 8.128e^{-11}$. Thus, 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 2.1 we observed that the GT can also be ambiguous and can have multiple interpretations. Thus, we performed a manual evaluation of the entities in the GT 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 2 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, F1 of 0.863 at a threshold of 0.5, also statistically significant.

Overall, the crowd improved the performance of the NER tools. In Figure 2 we see that the best performing threshold for Multi-NER+Crowd is on an interval of thresholds [0.5-0.7]. Furthermore, we see that the lower end of the interval is correlated with the best performing threshold for the crowd-driven ground truth (Multi-NER+CrowdGT). We believe this indicates that the entities in that interval are more prone to be ambiguous. Thus, *allowing for diversity provides better ground truth*.

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, has a significantly higher coverage of entities than any Single-NER tool. Second, through data inspection of the GT and the factors that answer for the increased number of FP and FN entities, we observed that the NER performance highly dependents on the ambiguity of such GT datasets. Third, our evaluation has shown that the crowd, *by harnessing the inter-annotator disagreement*, can correct the mistakes of the NER tools by reducing the number of FPs. Furthermore, the crowd-driven GT, that harnesses diversity, perspectives and granularities, proves to be a more reliable way of creating a GT when dealing with the natural language ambiguity and the overall task ambiguity.

[3]www.crowdflower.com

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

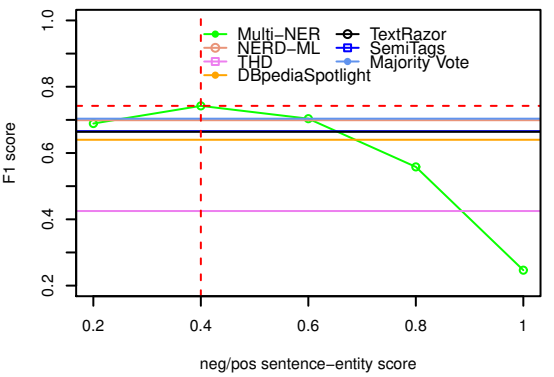

Fig. 1: Annotation quality F1 per negative/positive sentence-entity threshold

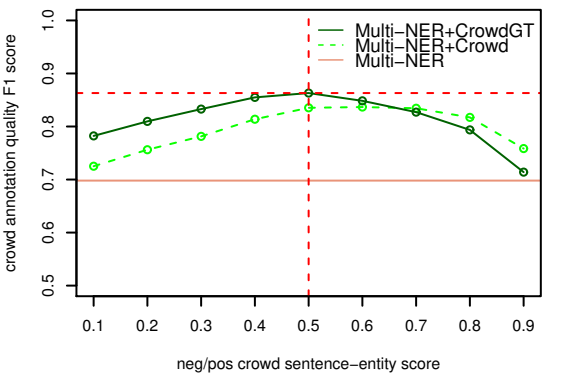

Fig. 2: Annotation quality F1 per negative/positive crowd sentence-entity threshold

Table II. : Single-NER vs. Multi-NER performance evaluation on entity surface and entity type

	TP					FP					FN				
	Place	People	Org	Role	Total	Place	People	Org	Role	Total	Place	People	Org	Role	Total
NERD-ML	130	189	177	116	612	23	24	72	23	142	34	220	67	73	394
SemiTags	136	225	167	1	529	21	4	26	3	54	28	184	77	188	477
THD	98	47	88	90	323	20	18	117	43	198	66	362	156	99	683
DBpedia-Spotlight	137	226	137	141	641	31	69	217	40	357	27	183	107	48	365
TextRazor	146	231	192	71	641	46	18	197	36	297	17	178	52	118	365
Multi-NER	161	297	230	169	857	75	104	371	100	650	4	112	14	20	149

REFERENCES

Lora Aroyo and Chris Welty. 2014. The Three Sides of CrowdTruth. *Journal of Human Computation* 1 (2014), 31–34. Issue 1. DOI:http://dx.doi.org/10.15346/hc.v1i1.3

Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: CrowdTruth and 7 Myths about Human Computation. *AI Magazine* 36, 1 (2015).

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics* 37, 4 (2011), 699–725.

Qiong Bu, Elena Simperl, Sergej Zerr, and Yunjia Li. 2016. Using microtasks to crowdsource DBpedia entity classification: A study in workflow design. *Semantic Web Journal* (2016).

Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2012. ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st international conference on WWW*. ACM, 469–478.

Oluwaseyi Feyisetan, Markus Luczak-Roesch, Elena Simperl, Ramine Tinati, and Nigel Shadbolt. 2015. Towards hybrid NER: a study of content and crowdsourcing-related performance factors. In *European Semantic Web Conference*. Springer, 525–540.

Aldo Gangemi. 2013. A comparison of knowledge extraction tools for the semantic web. In *Extended Semantic Web Conference*. Springer, 351–366.

Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. 2014. CrowdTruth: machine-human computation framework for harnessing disagreement in gathering annotated data. In *The Semantic Web–ISWC 2014*. Springer, 486–504.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines.. In *LREC*. 859–866.

Robert Voyer, Valerie Nygaard, Will Fitzgerald, and Hannah Copperman. 2010. A hybrid model for annotating named entity training corpora. In *Proceedings of the fourth linguistic annotation workshop*. ACL, 243–246.