Making Sense of Microposts (#Microposts2015) Named Entity rEcognition & Linking Challenge

Giuseppe Rizzo
EURECOM, France
giuseppe.rizzo@eurecom.fr

Amparo E. Cano
KMi, The Open University, UK
amparo.cano@open.ac.uk

Bianca Pereira
Insight Centre for Data Analytics, Ireland
bianca.pereira@insight-centre.org

Andrea Varga
Swiss Re, London, UK
varga.andy@gmail.com

ABSTRACT

Microposts are small fragments of social media content and a popular medium for sharing facts, opinions and emotions. Collectively, they comprise a wealth of data that is increasing exponentially, and which therefore presents new challenges for the Information Extraction community, among others. This paper describes the *Making Sense of Microposts* (#Microposts2015) Workshop's **Named Entity rEcognition and Linking (NEEL)** Challenge, held as part of the 2015 World Wide Web conference (WWW'15). The challenge task comprised automatic recognition and linking of entities appearing in different event streams of English Microposts on Twitter. Participants were set the task of investigating novel strategies for extracting entities in a tweet stream, typing these based on a set of pre-defined classes, and linking to DBpedia or NIL referents. They were also asked to implement a web service to run their systems, to minimize human involvement in the evaluation and allow measuring of processing times. The challenge attracted a lot of interest: 29 research groups expressed an intent to participate, out of which 21 signed the agreement required to be given a copy of the training and development datasets. Seven teams participated in the final evaluation of the challenge task, out of which six completed all requirements, including submission of an abstract describing their approach. The submissions covered sequential and joint linguistic methods, end-to-end and hybrid end-to-end, and linguistic approaches for tackling the challenge task. We describe the evaluation process and discuss the performance of the different approaches to the #Microposts2015 NEEL Challenge. We also release, with this paper, the #Microposts2015 NEEL Challenge Gold Standard, comprising the set of manually annotated tweets.

Keywords

Microposts, Named Entity Recognition, Named Entity Linking, Disambiguation, Knowledge Base, Evaluation, Challenge

1. INTRODUCTION

Microposts are short text messages published using minimal effort via social media platforms. They provide a publicly accessible wealth of data which has proven to be useful in different applications and contexts (e.g., music recommendation, social bots, spam detection, emergency response). However, extracting data from Microposts and linking it to external sources presents various challenges, due, among others, to the inherent characteristics of this type of data:

i) the restricted length;
ii) the noisy lexical nature, where terminology differs between users when referring to the same thing, and non-standard abbreviations are common.

A commonly used approach for making sense of Microposts is the use of textual cues, which provide contextual features for the underlying tweet content. One example of such a cue is the use of *Named Entities*. Extracting named entities from Microposts has, however, proven to be a challenging task; this was the focus of the Concept Extraction (CE) Challenge, part of the 2013 workshop, #MSM2013 [4]. A step further into the use of such cues is to ground entities in tweets by linking them to Knowledge Base referents. This prompted the Named Entity Extraction and Linking (NEEL) Challenge the following year, in #Microposts2014 [3]. These two research avenues, which add to the intrinsic complexity of the tasks proposed in 2013 and '14, prompted the Named Entity rEcognition and Linking (NEEL) Challenge in #Microposts2015. In NEEL 2015 we investigated further the role of the named entity type in the process, and the identification of named entities that cannot be grounded because they do not have a Knowledge Base referent. The English DBpedia 2014[1] dataset was the designated reference Knowledge Base for the 2015 NEEL challenge.

From the first Concept Extraction challenge (in 2013) through to the 2015 NEEL challenge, we have received over 40 submissions proposing state of the art approaches for extracting, typing, linking, and clustering relevant pieces of data from Microposts, namely, named entities. The purpose of each challenge was to set up an open and competitive environment that would encourage participants to deliver novel or improve on existing approaches for recognizing and linking entities from Microposts to either a reference Knowledge Base entry or NIL where such a reference does not exist. To encourage competition we solicited sponsorship for the winning submission, an award of €1,500. This was provided by SpazioDati,[2] a startup operating in the Big Data & Semantic Web market, who are active in the research community of entity linking.

[1] http://wiki.dbpedia.org
[2] http://www.spaziodati.eu

This generous sponsorship is testament to the growing interest in challenges related to automatic approaches for gleaning information from (the very large amounts of) social media data generated across all aspects of life, and whose knowledge content is recognised to be of value to industry.

This paper describes the #Microposts2015 NEEL Challenge, detailing its rationale and research challenges, the collaborative annotation of the corpus of Microposts, and our evaluation of the performance of each submission. We describe the approaches taken in the participants' systems – which use both established and novel, alternative approaches to entity extraction, typing, linking and clustering. The resulting body of work has implications for researchers, application designers and social media engineers who wish to harvest information from Microposts for their own objectives.

2. TASK DEFINITION AND EVALUATION

In this section we describe the goal of the challenge, the task set, and the process we followed to generate the corpus of Microposts.

2.1 The Task and Research Challenges

The 2015 challenge required participants to build automated systems to solve three main tasks:

i) extraction and typing of entity mentions within a tweet;
ii) linking of each mention to a referent in the English DBpedia 2014 dataset representing the same real world entity, or NIL for cases where no such entry exists;
iii) clustering of each unique, non-linked entity to a NIL identifier, where each cluster contains only mentions to the same real world entity.

In the rest of this paper we refer to the term appearing in a text as either an *entity mention* or simply an *entity*, while we refer to its DBpedia referent as the *candidate*. Consequently, the operation of entity detection is also referred to as *mention detection*, whilst for entity linking we use *candidate selection*.

An entity, in the context of this challenge, is used in the general sense of being, not requiring a material existence but only to be an instance of a taxonomy class. Thus, a mention of an entity in a tweet can be seen as a proper noun or an acronym. The extent of an entity is the entire string representing the name, excluding the preceding definite article (i.e., "the") and any other pre-posed (e.g., "Dr.", "Mr.") or post-posed modifiers.

In this task we consider an entity to be referenced in a tweet as a proper noun or an acronym when: i) it belongs to one of the categories specified in the NEEL Taxonomy (see Appendix A); and ii) it can be linked to an English DBpedia referent or to a NIL reference given the context of the tweet.

Pronouns (e.g., he/she, him/her) are not considered mentions of entities in the context of this challenge. Lowercase and compressed words (e.g., "c u 2night" rather than "see you tonight") are common in tweets. Thus, they are still considered mentions if they can be directly mapped to proper nouns. Complete entity extents, and not their substrings, are considered a valid mention. For example, from the following text excerpt: "Barack Obama gives a speech at NATO", neither of the words *Barack* nor *Obama* is considered by themselves, but rather *Barack Obama*. This is because they constitute a substring of the full mention [`Barack Obama`]. However, in the text: "Barack was born in the city, at which time his parents named him Obama" each of the terms [`Barack`] and [`Obama`] should be selected as a separate entity mention.

Nested entities with qualifiers should be considered as independent entities; similarly, compound entities should be annotated in isolation. E.g.,

Tweet:
```
Alabama CF Taylor Dugas has decided to
end negotiations with the Cubs and will
return to Alabama for his senior season.
#bamabaseball
```

For this tweet, the [`Alabama CF`] entity qualifies [`Taylor Dugas`]; the annotation for such a case should be: [`Alabama CF, Organization, dbp:Alabama_Crimson_Tide`] and [`Taylor Dugas, Person, NIL1`], where NIL1[3] is the unique NIL identifier describing the real world entity "Taylor Dugas".

2.1.1 *Noun phrases completing the definition of an entity*

In the 2015 challenge, as opposed to the previous edition, not all noun phrases are considered as entity mentions. E.g., in:

Tweet:
```
I am happy that an #asian team have
won the womens world cup!  After just
returning from #asia i have seen how
special you all are!  Congrats
```

While "asian team" could be considered as an Organization-type it can refer to multiple entities. Therefore we do not consider it as an entity mention, and it should not be annotated.

While noun phrases can be linked to existing entities, we do not consider them as entity mentions. In such cases we only keep "embedded" entity mentions. E.g., in:

Tweet:
```
head of sharm el sheikh hospital is
DENYING
```

"head of sharm el sheikh hospital" refers to a Person-type; however, since it is not a proper noun we do not consider it as an entity mention. For that reason, in this case the annotation should only contain the embedded entity [`sharm el sheikh hospital`]: [`sharm el sheikh hospital, Organization, dbp: Sharm_International_Hospital`].

In the tweet:

[3] NIL1 is composed of two parts: NIL and the suffix 1. Any suffix, numeric or alphanumeric, is considered as a valid suffix.

Tweet:
The best Panasonic LUMIX digital camera from a wide range of models

while digital camera describes the entity "Panasonic LUMIX", it is not considered within the entity annotation, since it is used in the context as a noun phrase.[4] In this case the annotation should be [Panasonic, ORG, dbp:Panasonic][LUMIX, Product, dbp:Lumix].

Entity mentions in a tweet can also be typified based on the context in which they are used. In:

Tweet:
Five New Apple Retail Stores Opening Around the World: As we reported, Apple is opening 5 new retail stores on ...

In this case [Apple Retail Stores] refers to a Location-type, while the second [Apple] mention refers to an Organisation-type.

2.1.2 *Special Cases in Social Media (# and @)*

Entities may be referenced in a tweet preceded or composed by # and @, e.g.:

Tweets:
#[Obama] is proud to support the Respect for Marriage Act.
#[Barack Obama] is proud to support the Respect for Marriage Act.
@[BarackObama] is proud to support the Respect for Marriage Act.

Hashtags (i.e., words referenced by a #) can refer to entities, but this does not mean that all hashtags will be considered as entities. Further, for our purposes, the characters # and @ should not be included in the annotation string. We consider the following cases:

Hashtagged nouns and noun-phrases:

Tweet:
I burned the cake again. #fail

The hashtag "#fail" does not represent an entity. Thus, it should not be annotated as an entity mention.

Partially tagged entities:

Tweet:
Congrats to Wayne Gretzky, his son Trevor has officially signed with the Chicago @Cubs today

Here "Chicago @Cubs" refers to the proper noun characterising the [Chicago Cubs] entity. (Note that in this case "Chicago" is not a qualifier, but rather, part of the entity mention.) The annotation should therefore be [Chicago, Organization, dbp:Chicago_Cubs] and [Cubs, Organization, dbp:Chicago_Cubs].

Tagged entities:

If a proper noun is split and tagged with two hashtags, the entity mention should be split into two separate mentions.

Tweet:
#Amy #Winehouse

In this case we annotate [Amy, Person, dbp:Amy_Winehouse] [Winehouse, Person, dbp:Amy_Winehouse]

2.1.3 *Use of Nicknames*

The use of nicknames (i.e., descriptive names replacing the actual name of an entity) are commonplace in Social Media, e.g., the use of "SFGiants" to refer to "the San Francisco Giants". For these cases, nicknames are co-referenced to the entity they refer to in the context of a tweet.

Tweet:
#[Panda] with 3 straight hits to give #[SFGiants] 6-1 lead in 12th

We annotate [Panda, Person, dbp:Pablo_Sandoval] and [SFGiants, Organization, dbp:San_Francisco_Giants].

2.2 Evaluation Strategy

Participants were required to implement their systems as a publicly accessible web service following a REST-based protocol, in order to submit (up to 10) contending entries to a registry of the NEEL challenge services. In this context, we refer to a contending entry as the participant's REST endpoint queried in the evaluation campaign. Each endpoint had a Web address (URI) and a name, which we defined as run_{ID}. Upon receiving the registration of the REST endpoint, calls to the contending entry were scheduled in two different time windows, namely, D-Time – to test the APIs, and T-Time – for the final evaluation and metric computations. To ensure correctness of the results and avoid any loss we triggered a large number of queries and statistically evaluated the results.

2.2.1 *Metrics and Scorer*

The evaluation was conducted using four different metrics:

[4] Panasonic LUMIX refers to a series of cameras. Therefore to be considered a proper noun it should be followed by a number or an identifier.

i) strong_typed_mention_match,
ii) strong_link_match,
iii) mention_ceaf,
iv) latency.

The *strong_typed_mention_match* evaluates the micro average F_1 score for all annotations considering the mention boundaries and their types. The *strong_link_match* is the micro average F_1 score for annotations considering the correct link for each mention. The *mention_ceaf* (Constrained Entity-Alignment F-measure) [10] is a clustering metric developed to evaluate clusters of annotations. It evaluates the F_1 score for both NIL and non-NIL annotations in a set of mentions. The *latency* measures the computation time of an entry (in seconds), to annotate a tweet. The final score is computed according to Equation 1. The *latency* metric was included only to resolve cases where there was a tie in the evaluation score.

$$\begin{aligned} score = {} & 0.4 * mention\_ceaf \\ & + 0.3 * strong\_typed\_mention\_match \\ & + 0.3 * strong\_link\_match \end{aligned} \tag{1}$$

The scorer proposed for the TAC KBP 2014 task[5] was used to perform the evaluation.

2.2.2 Selection of the Annotation Results

Algorithm 1 EVALUATE($E, Tweet, N = 100, M = 30$)

```
for all $e_i \in E$ do
    $A^S = \emptyset, L^S = \emptyset$
    for all $t_j \in Tweet$ do
        for all $n_k \in N$ do
            $(A, L) = annotate(t_j, e_i)$
        end for

        // Majority Voting Selection of $a$ from $A$
        for all $a_k \in A$ do
            $hash(a_k)$
        end for
        $A_j^S$ = Majority Voting on the exact same $hash(a_k)$

        // Random Selection of $l$ from $L$
        generate $L^T$ from the uniformly random selection of M $l$ from $L$
        $(\mu, \sigma)$ = computeMuAndSigma($L^T$)
        $L_j^S$ = $(\mu, \sigma)$
    end for
end for
```

To ensure the correctness of the results and avoid any loss we triggered N (with N=100) calls to each entry. We then applied a majority voting approach over the set of annotations per tweet and statistically evaluated the latency by applying the law of large numbers [14]. Algorithm 1 provides a sketch of the algorithm used during the evaluation campaign.

3. PARTICIPANT OVERVIEW

The challenge attracted a lot of interest from research groups spread around the world. Twenty-nine groups expressed their intent to participate in the challenge; out of which twenty-one signed the agreement required to be given a copy of the training and development datasets. Seven teams participated in the final evaluation of the challenge task, out of which six completed submission with an abstract describing the approach they took. The final submissions are listed in Table 1.

Table 2 provides a taxonomy of the approaches proposed this year for tackling the challenge task. From an historical perspective, starting from the first Concept Extraction (CE) challenge till the current, 2015, apart from the NIL detection and clustering introduced in this challenge, we observed:

1. the consolidation of a normalization procedure, namely preprocessing, to increase the expressiveness of the tweets, e.g. via expansion of Twitter accounts and hashtags with the actual names of entities they represent;
2. the consolidated contribution of Knowledge Bases in the Mention Detection and Typing task. This leads to higher coverage, which, along with the linguistic analysis and type prediction, better fits the Microposts domain;
3. the consolidation of the Candidate Selection performed as an End-to-End approach. Such an approach has been further developed with the addition of fuzzy distance functions operating over n-grams and acronyms;
4. a considerable decrease in off-the-shelf systems.

We provide next a detailed description of each contribution.

In [15], Yamada et al., present a five-sequential stage approach: preprocessing, generation of potential entity mentions, candidate selection, NIL detection, and entity mention typing. In the preprocessing stage, they propose a tokenization and Part-of-Speech (POS) tagging approach based on [7], along with the extraction of tweet timestamps. They tackle the generation of potential entity mentions by computing n-grams (with $n = 1..10$ words) and matching them to Wikipedia titles, Wikipedia titles of the redirect pages, and anchor text using exact, fuzzy, and approximate match functions. An in-house dictionary of acronyms is built by splitting the mention surface into different n-grams (where 1 n-gram corresponds to 1 char). At this stage all entity mentions are linked to their candidates, i.e., the Wikipedia counterparts. The candidate selection is approached as a learning to rank problem: to each mention is assigned a confidence score computed as the output of a supervised learning approach using Random Forest as the classifier. An empirically defined threshold is used to select the relevant mentions; in the case of mention overlap the span with the highest score is selected. The NIL detection is tackled as a supervised learning task, in which Random Forest is used. The features used are the predicted entity types, contextual features such as surrounding words, POS, length of the n-gram and capitalization features. The mention entity typing stage is treated as a supervised learning task where two independent classifiers are built: a Logistic Regression classifier for typing entity mentions and a Random Forest for typing NIL entries.

Gârbacea et al., [6] present a sequential approach composed of four stages: entity mention detection, candidate selection, NIL clustering, and resolution of overlapping mentions. The first stage is tackled by empowering both an annotation-based off-the-shelf system, Semanticizer,[6] and a Named Entity Recognition classifier trained using the challenge dataset. For each entity mention, a Learning to Rank supervised model is used to select the most representative DBpedia reference of the entity mention (candidate detection). The resulting type of the DBPedia reference entity is used to type the

[5] `https://github.com/wikilinks/neleval/wiki/Evaluation`

[6] `https://github.com/semanticize/semanticizer`

Table 1: Accepted submissions with team affiliations and number of runs for each.

Reference	Team's affiliation	Team Name	Authors	No. of entries
[15]	Studio Ousia and Keio University and National Institute of Informatics	ousia	Yamada *et al.*	10
[6]	University of Amsterdam	uva	Gârbacea *et al.*	10
[2]	University of Bari	uniba	Basile *et al.*	2
[8]	University of Alberta	ualberta	Guo *et al.*	1
[9]	Amrita Vishwa Vidyapeetham	cen_neel	Barathi Ganesh *et al.*	1
[13]	IIT Kharagpur	tcs-iitkgp	Sinha *et al.*	3

Table 2: Overview summary of approaches applied in the #Microposts2015 NEEL Challenge.

Step	Method	Features	Knowledge Base	Off-the-Shelf Systems
Preprocessing	Cleaning Expansion Extraction	stop words, spelling dictionary, acronyms, hashtags, Twitter accounts, tweet timestamps, punctuation, capitalization, token positions		
Entity Mention Detection	Approximate String Matching, Exact String Matching, Fuzzy String Matching, Acronym Search Perfect String Matching, Levenshtein Matching, Jaccard String Matching, Prior Probability Matching, Context Similarity Matching, Conditional Random Fields, Random Forest	POS, tokens and adjacent tokens, contextual features, tweet timestamps, string similarity, n-grams, proper nouns, mention similarity score, Wikipedia titles, Wikipedia redirects, Wikipedia anchors, word embeddings	Wikipedia, DBpedia	Semanticizer
Entity Typing	DBpedia Type, Logistic Regression, Random Forest, Conditional Random Fields	tokens, linguistic features, word embeddings, entity mentions, NIL mentions DBpedia and Freebase types	DBpedia Freebase	
Candidate Selection	Distributional Semantic Model, Random Forest, RankSVM, Random Walk with Restart, Learning to Rank	gloss, contextual features, graph distance	Wikipedia, DBpedia	DBpedia Spotlight
NIL Detection	Conditional Random Fields, Random Forest, Lack of candidate, Score Threshold	POS, contextual words, n-grams length, predicted entity types, capitalization ratio		
NIL Clustering	Surface Form Aggregation, Type Aggregation	entity mention label, entity mention type		

entity mention (the normalization of the type is performed via a manual alignment from the DBpedia ontology and the NEEL taxonomy). The NIL is finally solved using a clustering algorithm operating on the lexical similarity of the entity mentions that do not have any DBpedia referents. To resolve the entity mention overlaps, they create a graph of all non-overlapping mentions, and assign a link score (non-linked mentions get a fixed score). They then find the highest scoring path through the graph using dynamic programming, and return the mentions of this path as the resolved list of mentions.

The system presented in Basile et al. [2] also follows a sequential workflow of mention detection and candidate selection. For the former, two approaches are built: an unsupervised based on the extraction of n-grams ($n = 0..5$), and a supervised based on the prediction of the entity boundaries from a POS tagger. Each potential entity mention is then matched with a list of DBpedia concept titles using the Levenshtein Distance, Jaccard Index, and Lucene similarity output. A filter of the entity mentions is applied with a similarity threshold of 0.85. The candidate selection stage then resolves the ambiguity of the several potential links identifying an entity mention through an adaptation of the distributional Lesk algorithm [1]. Finally, entity typing is carried out by inheriting the DBpedia type of the DBpedia reference entity pointed to, and then manually aligning this to the NEEL taxonomy.

In [8], Guo et al., present a sequential approach to the NEEL task. First, they generate potential entity mentions, using TwitIE. They then link those mentions to corresponding DBpedia referents via a candidate selection algorithm based on the similarity of the text to a dictionary built from Wikipedia titles, redirect pages, disambiguation pages and anchor text. Mentions that are not linkable are flagged as NIL. The problem of finding the correct candidate to be linked to each mention is tackled using Random Walks. Starting from the candidate links retrieved from DBpedia, a subgraph of DBpedia is built adding all adjacent entity mentions to the candidates. A personalized PageRank is then executed, giving more importance to unambiguous entities. Finally, measures of semantic relatedness between entity links, prior probability and context similarity are combined to compute an overall score. The candidate with the highest score is considered as the correct link. NIL clustering uses string similarity of entity mention names.

In [9], Barathi et al., present another sequential pipeline to the 2015 challenge, composed of generation of potential entity mentions, mention detection and candidate selection. The first stage is tackled with a linguistic approach that tokenizes the text according to Twitter cues, such as hashtags and emoticons, using the TwitIE tagger. The system then classifies entity mentions by applying a supervised learning approach using direct (e.g., POS tags) and indirect features (two words on the left and right of a candidate mention entity). In total, the authors use 34 lexical features and experiment with 3 different supervised learning algorithms. The final system implements what is determined to be the best entity recognition configuration, based on the performance achieved in the development test. The candidate selection stage is tackled by looking up DBpedia referent links. The candidate link which maximizes the similarity score between related entries and the mentions is designated as the representative. Entity mentions without related links are assigned to NIL.

Sinha et al., [13] also follow a sequential approach to the challenge task, by first detecting entity mentions from the text, and then selecting the most representative DBpedia referents (candidate selection). The first stage grounds on the linguistic cues extracted from conventional linguistic approaches such as POS tagging, word capitalization, and hashtag in the tweet. A Conditional Random Field (CRF) classifier is then trained with the linguistic features and the contextual similarity of adjacent tokens, with token window set to 5. The candidate selection is performed using an entity resolution mechanism that takes as input both the output of the entity mention detection stage and the output of DBpedia Spotlight [5]. For each entity returned from DBpedia Spotlight, if (i) the retrieved entity is found to be a substring of any of the extracted mentions in the entity mention detection stage, and if (ii) a substring match is found, then the corresponding DBpedia referent is returned and assigned to the final entity mention. If there is no match to the mention entities being extracted by the entity mention detection stage and those extracted by DBpedia Spotlight, they are assigned as NIL.

4. CORPUS CREATION AND ANNOTATION

In this section we describe the challenge dataset and the annotation process for characterising it and generating the Gold Standard. Since the challenge task was to automatically recognise, type, and link named entities (either to DBpedia referents or NIL identifiers), we built the challenge dataset considering both event and non-event tweets. While event tweets are more likely to contain named entities, non-event tweets enable us to evaluate system performance in avoiding false positives in the mention detection and candidate selection stages. The challenge dataset comprises tweets from the years 2011, 2013 and 2014. Tweets from 2011 and 2013 were extracted from a collection of over 18 million tweets provided by the Redites project.[7] These tweets cover multiple noteworthy events from 2011 and 2013 (including the death of Amy Winehouse, the London Riots, the Oslo bombing and the Westgate Shopping Mall terrorist attack). To obtain a dataset containing both event and non-event tweets, we also collected tweets from the Twitter firehose in November 2014 covering both event (such as the UCI Cyclo-cross World Cup) and non-event tweets.

Table 3: General statistics of the #Microposts2015 NEEL corpus. Dev refers to the Development set, while NEs refers to Named Entities.

	Training	Dev	Test
No. of Tweets	3,498	500	2,027
No. of Words	13,752	3,281	10,274
No. of Tokens	67,393	7,845	35,558
Avg. Tokens/Tweet	19.27	15.69	17.54
No. of Tweets with NEs	2,023	387	1,663
No. of NEs	4,016	790	3,860
No. of NIL NEs	451	362	1,478
No. of NEs with Referents	3,565	428	2,382
Avg. NEs/Tweet	1.985	2.041	2.321
Avg. NIL NEs/Tweet	0.222	0.935	0.888
Avg. NEs with Referents/Tweet	1.762	1.105	1.432

4.1 Corpus Description

The corpus consists of three main datasets: Training (58%), Development (8%) – which enabled participants to tune their systems – and Test (34%). The statistics describing the data are provided in Table 3.[8] The Training set comprises 3,498 tweets, with 67,393 tokens and 4,016 named entities. This dataset corresponds to the entire corpus of the #Microposts2014 NEEL Challenge[9] (Training + Test sets), extended with annotations for additional entity types (including Character, Event, Product, Thing) and NIL references. We also harmonized the candidate selection with the rigid designation of entity in this challenge. The Development dataset consists of 500 tweets, with 7,845 tokens and 790 named entities, while the Test set contains 35,558 tokens and 3,860 named entities. These two datasets were created by excluding the #Microposts2014 NEEL tweets from the 2015 challenge dataset, and randomly splitting the remaining tweets. The Training dataset presented a higher rate of named entities linked to DBpedia (88.76%), while the Development and Test sets were more challenging, presenting only 54.18% and 61.71% respectively. The percentage of tweets mentioning at least one entity is 57.83% in the Training set, 77.4% in the Development (Dev) set, and 82.05% in the Test set. There is very little overlap of named entities between the Training and Test data, with 4.6% (186) of the named entities in the Training also occurring in the Test set.

Summary statistics of the entity types are provided in Table 4. Across the 3 datasets the most frequent types are Person, Organization and

[7] http://demeter.inf.ed.ac.uk/redites

[8] For the computation of the statistics, the tweets were tokenized using the TwitterNLP tool (http://www.ark.cs.cmu.edu/TweetNLP).

[9] http://ceur-ws.org/Vol-1141/microposts2014-neel_challenge_gs.zip

Location. The Training dataset presents a higher rate of Organization and Thing types on average, compared to the Dev and Test datasets. The Dev dataset presents a higher rate of named entities mentioning events. The Test dataset presents a higher rate of Location. Product-types are distributed nearly evenly across the three datasets. The distributional differences between the entity types in the three sets can be clearly seen. This makes the #Microposts2015 NEEL task challenging, particularly when tackled with supervised learning approaches.

Table 4: Entity type statistics for the three data sets. Dev refers to the Development set.

Type	Training	Dev	Test
Character	43 (1.07%)	5 (0.63%)	15 (0.39%)
Event	182 (4.53%)	81 (10.25%)	219 (5.67%)
Location	786 (19.57%)	132 (16.71%)	957 (24.79%)
Organization	968 (24.10%)	125(15.82%)	541 (14.02%)
Person	1102 (27.44%)	342 (43.29%)	1402 (36.32%)
Product	541 (13.47%)	80 (10.13%)	575 (14.9%)
Thing	394 (9.81%)	25 (3.16%)	151 (3.92%)

4.2 Generating the Gold Standard

The Gold Standard (GS) was generated with the help of 3 annotators. The annotation process followed six stages.

Stage 1. Unsupervised annotation of the corpus was performed, to extract the potential entity mentions, along with the corresponding entity types and candidate links to DBpedia, that were used as input to the next stage. At this stage we used the system described in [12] for annotation.

Stage 2. The data set was divided into 3 batches (Training, Development, Test). Two annotators, using GATE,[10] annotated each batch. GATE was selected because the annotation process is guided by an ontology-centric view. However, we encountered a few issues adding the link property to each annotation, which slowed down the process, because of low flexibility in interaction with the user interface. A set of guidelines for annotation was also written, to guide the annotators in *i)* selecting the entity mentions, their types, and the corresponding candidate links provided in the first stage, and then *ii)* adding any missing annotation. The annotators were also asked to mark any problematic cases encountered.

Stage 3. A third annotator, knowledgeable about the protocol followed in Stages 1 and 2, went through the problematic cases and, involving the two initial annotators, refined the annotation procedures. The annotators then looped through stages 2 and 3 of the process till most problematic cases were resolved.

Stage 4. Unsupervised NIL Clustering generation, based on mention strings and their types, was performed.

Stage 5. The third annotator went through all NILs to include or exclude them from a given cluster. The number of mentions per NIL cluster is presented in Table 5. This shows that the Entity Type Event represented a tougher challenge for the NIL Clustering ,while the other Types had, on average, number of mentions very close to one.

Stage 6. the so-called *Adjudication Stage*, where the challenge participants reported incorrect or missing annotations. Each reported mention was evaluated by one of the challenge chairs to check compliance with the Challenge Annotation Guidelines, and additions and corrections made as required.

Table 5: Average number of mentions per NIL Cluster for each Named Entity type.

Type	Training	Dev	Test
Character	1.50	1.00	1.00
Event	1.67	4.50	6.11
Location	1.00	1.00	1.20
Organization	1.52	1.08	1.24
Person	1.12	1.16	1.50
Product	1.96	1.03	1.36
Thing	1.00	1.00	1.00

5. CHALLENGE RANKING

Table 6 provides the #Microposts2015 NEEL rankings. As a baseline we used a state-of-the-art approach for recognizing and linking entities from short text that is developed by *acubelab*. The system is described in [11]. The ranking is based on Equation 1, which linearly weights the contribution of the 3 metrics used in the evaluation, measuring, respectively, the contribution of the clustering approach (mention_ceaf), the typing component (strong_typed_-mention_match) and the linking stage (strong_link_match). Team *ousia* [15] outperformed all other participants, with a 69% performance increase with respect to the second ranked approach, the baseline system. The top-ranked approach in this noisy context underlines current and ongoing research and industrial path in pushing toward an End-to-End system, augmented by the linguistic strength of a conventional pipeline used to filter out the irrelevant entity mentions. This approach recasts the NIL clustering stage and a supervised learning approach in predicting the role and the type of named entities that are not yet available in a Knowledge Base, such as emergent named entities, or named entities not in the scope of the Knowledge Base.

The Annotation results for the group tcs-iitkgp [13] were excluded from the ranking as they were not compatible with the challenge guidelines.

Table 6: Final #Microposts2015 NEEL Ranking

Rank	Reference	Team Name	run_{ID}	r_S
1	[15]	ousia	9	0.8067
2	**[11]**	**acubelab**	**7**	**0.4757**
3	[6]	uva	2	0.4756
4	[2]	uniba	uniba-sup	0.4329
5	[8]	ualberta	ualberta	0.3808
6	[9]	cen_neel	cen_neel_1	0.0004

Table 7 details the performance according to the *metric mention_ceaf* of the top ranked run for each participant. The runs are sorted according to the F_1 measure.

[10] https://gate.ac.uk

Table 7: Breakdown mention_ceaf figures per participant.

Rank	Reference	Team Name	run_{ID}	F_1
1	[15]	ousia	9	0.84
2	[6]	uva	2	0.643
3	**[11]**	**acubelab**	**7**	**0.506**
4	[2]	uniba	uniba-sup	0.459
5	[8]	ualberta	ualberta	0.394
6	[9]	cen_neel	cen_neel_1	0.001

Table 8 reports the performance of the top ranked run per participant according to the metric *strong_typed_mention_match*. The runs are sorted according to the F_1 measure.

Table 8: Breakdown strong_typed_mention_match figures per participant.

Rank	Reference	Team Name	run_{ID}	F_1
1	[15]	ousia	9	0.807
2	[6]	uva	2	0.412
3	**[11]**	**acubelab**	**7**	**0.388**
4	[2]	uniba	uniba-sup	0.367
5	[8]	ualberta	ualberta	0.329
6	[9]	cen_neel	cen_neel_1	0

Table 9 reports the performance of the top ranked run per participant according to the metric *strong_link_match*. The runs are sorted according to the F_1 measure.

Table 9: Breakdown strong_link_match figures per participant.

Rank	Reference	Team Name	run_{ID}	F_1
1	[15]	ousia	9	0.0.762
2	**[11]**	**acubelab**	**7**	**0.523**
3	[2]	uniba	uniba-sup	0.464
4	[8]	ualberta	ualberta	0.415
5	[6]	uva	2	0.316
6	[9]	cen_neel	cen_neel_1	0

Table 10 reports the performance of the top ranked run per participant based on latency (expressed in seconds s). Each measure is reported along with the confidence interval obtained from the selection procedure of the annotation results as reported in 2.2.2.

Finally, Table 11 shows the breakdown for the best 3 runs per participant over all metrics used in the evaluation of the systems.

6. CONCLUSIONS

The #Microposts2014 NEEL challenge was to foster the development of novel approaches for entity extraction, and linking in Microposts. In 2015 the NEEL task was extended to include integration of named entity typing and the characterization of entities to either DBpedia referents or NIL references. The motivation for organizing this challenge is the strong, current interest of the research and commercial communities in developing systems able to fit the challenging context of Microposts in entity extraction, entity recognition, and entity linking. Although state-of-the-art approaches offer a large number of options for tackling the challenge task, the evaluation results show that the NEEL task remains challenging when applied to tweets with their peculiarities, compared to standard, lengthy texts.

Table 10: Breakdown latency figures per participant.

Rank	Reference	Team Name	run_{ID}	$[s]$
1	**[11]**	**acubelab**	**7**	**0.13±0.02**
2	[6]	uva	2	0.19±0.09
3	[2]	uniba	uniba-sup	2.03±2.35
4	[8]	ualberta	ualberta	3.41±7.62
3	[15]	ousia	9	8.5±3.62
6	[9]	cen_neel	cen_neel_1	12.37±27.6

The evaluation strategy used in the 2014 challenge has been extended in 2015, to account for *mention_ceaf*, *strong_link_match*, *strong_typed_mention_match* and *latency*, following the established metrics introduced in the TAC KBP 2014 task. Carrying out evaluation in this way provided a more robust approach for ranking participants' entries.

As a result of the 2015 NEEL challenge we have generated a manually annotated corpus, which extends that in 2014 with the annotation of typed entities and the generation of NIL identifiers. To the best of our knowledge this is the largest publicly available corpus providing named entities, types, and link annotations for Microposts. The gold standard[11] is released with the CC BY 4.0 license.[12] We hope that through our release of data and resources, we will promote research on entity recognition and disambiguation, especially with regard to Microposts.

Our evaluation results report a clear winner: Team *ousia* [15] consolidated and, further, extended the findings of the NEEL 2014 winner, using an End-to-End system for both candidate selection and mention typing, along with a linguistic pipeline to perform entity typing and filtering.

The #Microposts2015 NEEL challenge saw a considerable drop in participants after the initial intent to participate. Among the participants who withdrew, reasons given were mainly poor results from their prototypes and the complexity in developing a reliable prototype to be deployed as a Web service. Aiming to consolidate the current challenge task we believe it will aid participants in such challenges to further develop their prototypes by providing a base engineering platform for deployment in a live context. We have, in 2015, also built bridges with the TAC community. We plan to strengthen these and to involve a larger audience of potential participants spanning the Linguistics, Machine Learning, Knowledge Extraction and Data Semantics fields, in order to widen the scope for potential solutions to what is acknowledged to be a challenging, albeit valuable, exercise.

7. ACKNOWLEDGMENTS

Especial thanks to SpazioDati who generously sponsored the prize for the winning submission. We thank also the participants who helped to improve the Gold Standard.
Giuseppe Rizzo is funded by LinkedTV (GA No. 287911), Amparo

[11] http://ceur-ws.org/Vol-1395/microposts2015_neel-challenge-report/microposts2015-neel_challenge_gs.zip

[12] http://creativecommons.org/licenses/by/4.0

Table 11: Top 3 runs per participant, sorted according to r_S.

Rank	Reference	Team Name	run_{ID}	$tagging_{F1}$	$clustering_{F1}$	$linking_{F1}$	$latency[s]$	r_S
1	[15]	ousia	9	0.807	0.84	0.762	8.5±3.62	0.8067
2	[15]	ousia	5	0.68	0.843	0.762	8.48±3.6	0.7698
3	[15]	ousia	10	0.679	0.842	0.762	8.49±3.57	0.7691
4	[11]	acubelab	7	0.388	0.506	0.523	0.13±0.02	0.4757
5	[6]	uva	2	0.412	0.643	0.316	0.19±0.09	0.4756
6	[11]	acubelab	6	0.385	0.506	0.524	0.13±0.02	0.4751
7	[11]	acubelab	9	0.388	0.506	0.523	0.13±0.02	0.4734
8	[6]	uva	3	0.404	0.642	0.285	0.19±0.1	0.4635
9	[6]	uva	6	0.383	0.595	0.318	1.73±0.86	0.4483
10	[2]	uniba	uniba-sup	0.367	0.459	0.464	2.03±2.35	0.4329
11	[8]	ualberta	ualberta	0.329	0.394	0.415	3.41±7.62	0.3808
12	[2]	uniba	uniba-unsup	0.367	0.459	0.464	2.03±2.35	0.4329
13	[9]	cen_neel	cen_neel_1	0	0.001	0	12.89±27.6	0.004

E. Cano is funded by the MK:smart project and Bianca Pereira by the Science Foundation Ireland (GA No. SFI/12/RC/2289).

8. REFERENCES

[1] P. Basile, A. Caputo, and G. Semeraro. An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model. In *25th International Conference on Computational Linguistics (COLING'14)*, 2014.

[2] P. Basile, A. Caputo, G. Semeraro, and F. Narducci. UNIBA: Exploiting a distributional semantic model for disambiguating and linking entities in tweets. In *5th Workshop on Making Sense of Microposts (#Microposts2015)*, pages 62–63, 2015.

[3] A. E. Cano Basave, G. Rizzo, A. Varga, M. Rowe, M. Stankovic, and A.-S. Dadzie. Making Sense of Microposts (#Microposts2014) Named Entity Extraction & Linking Challenge. In *4th Workshop on Making Sense of Microposts (#Microposts2014)*, 2014.

[4] A. E. Cano Basave, A. Varga, M. Rowe, M. Stankovic, and A.-S. Dadzie. Making Sense of Microposts (#msm2013) Concept Extraction Challenge. In *Making Sense of Microposts (#MSM2013) Concept Extraction Challenge*, 2013.

[5] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *9th International Conference on Semantic Systems (I-SEMANTICS '13)*, 2013.

[6] C. Gârbacea, D. Odijk, D. Graus, I. Sijaranamual, and M. de Rijke. Combining multiple signals for semanticizing tweets: University of Amsterdam at #Microposts2015. In *5th Workshop on Making Sense of Microposts (#Microposts2015)*, pages 59–60, 2015.

[7] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL'11)*, 2011.

[8] Z. Guo and D. Barbosa. Entity recognition and linking on tweets with random walks. In *5th Workshop on Making Sense of Microposts (#Microposts2015)*, pages 57–58, 2015.

[9] B. G. H B, A. N, A. K. M, V. R, and S. K P. AMRITA – CEN@NEEL: Identification and linking of Twitter entities. In *5th Workshop on Making Sense of Microposts (#Microposts2015)*, pages 64–65, 2015.

[10] X. Luo. On coreference resolution performance metrics. In *Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP'05)*, 2005.

[11] F. Piccinno and P. Ferragina. From TagME to WAT: A New Entity Annotator. In *1st International Workshop on Entity Recognition & Disambiguation (ERD '14)*, 2014.

[12] G. Rizzo, M. van Erp, and R. Troncy. Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web. In *9th International Conference on Language Resources and Evaluation (LREC'14)*, 2014.

[13] P. Sinha and B. Barik. Named entity extraction and linking in #Microposts. In *5th Workshop on Making Sense of Microposts (#Microposts2015)*, pages 66–67, 2015.

[14] R. Walpole and R. Myers. *Probability and statistics for engineers & scientists (Eighth Edition)*. Pearson Education International, 2007.

[15] I. Yamada, H. Takeda, and Y. Takefuji. An end-to-end entity linking approach for tweets. In *5th Workshop on Making Sense of Microposts (#Microposts2015)*, pages 55–56, 2015.

APPENDIX

A. NEEL TAXONOMY

`Thing`

- languages
- ethnic groups
- nationalities
- religions
- diseases
- sports
- astronomical objects

Examples:

`If all the #[Sagittarius] in the world`
`Jon Hamm is an [American] actor`

`Event`

- holidays
- sport events
- political events

social events

Examples:

```
[London Riots]
[2nd World War]
[Tour de France]
[Christmas]
[Thanksgiving] occurs the ...
```

`Character`

fictional characters

comic characters

title characters

Examples:

```
[Batman]
[Wolverine]
[Donald Draper]
[Harry Potter] is the strongest wizard in the school
```

`Location`

public places (squares, opera houses, museums, schools, markets, airports, stations, swimming pools, hospitals, sports facilities, youth centers, parks, town halls, theatres, cinemas, galleries, universities, churches, medical centers, parking lots, cemeteries)

regions (villages, towns, cities, provinces, countries, continents, dioceses, parishes) commercial places (pubs, restaurants, depots, hostels, hotels, industrial parks, nightclubs, music venues, bike shops)

buildings (houses, monasteries, creches, mills, army barracks, castles, retirement homes, towers, halls, rooms, vicarages, courtyards)

Examples:

```
[Miami]
Paul McCartney at [Yankee Stadium]
president of [united states]
Five New [Apple Retail Store] Opening Around
```

`Organization`

companies (press agencies, studios, banks, stock markets, manufacturers, cooperatives)

subdivisions of companies

brands

political parties

government bodies (ministries, councils, courts, political unions)

press names (magazines, newspapers, journals)

public organizations (schools, universities, charities)

collections of people (sport teams, associations, theater companies, religious orders, youth organizations, musical bands)

Examples:

```
[Apple] has updated Mac Os X
[Celtics] won against
[Police] intervene after disturbances
[Prism] performed in Washington
[US] has beaten the Japanese team
```

`Person`

people's names (titles and roles are not included, such as Dr. or President)

Examples:

```
[Barack Obama] is the current
[Jon Hamm] is an American actor
[Paul McCartney] at Yankee Stadium
call it [Lady Gaga]
```

`Product`

movies

tv series

music albums

press products (journals, newspapers, magazines, books, blogs)

devices (cars, vehicles, electronic devices)

operating systems

programming languages

Examples:

```
Apple has updated [Mac Os X]
Big crowd at the [Today Show]
[Harry Potter] has beaten any records
Washington's program [Prism]
```