Making Sense of Microposts (#Microposts2016) Named Entity rEcognition and Linking (NEEL) Challenge

Giuseppe Rizzo
ISMB, Italy
giuseppe.rizzo@ismb.it

Marieke van Erp
Vrije Universiteit Amsterdam,
The Netherlands
marieke.van.erp@vu.nl

Julien Plu
EURECOM, France
julien.plu@eurecom.fr

Raphaël Troncy
EURECOM, France
raphael.troncy@eurecom.fr

ABSTRACT

This paper describes the 2016 **Named Entity rEcognition and Linking (NEEL)** Challenge, held as track of the *Making Sense of Microposts* Workshop co-located with the *World Wide Web* conference (WWW). The challenge task comprised of automatic linking and classification of entities appearing in different event streams of English tweets. Participants were invited to develop novel strategies for extracting entities in a tweet stream, typing them based on a set of pre-defined classes, and linking them to resources from the DBpedia 2015-04 knowledge base or NIL referents. The challenge attracted a lot of interest: 37 research teams expressed an intent to participate and signed the agreement to acquire the dataset. Six different approaches took part to the final evaluation of the challenge task. The submissions covered joint linguistic and graph-based entity recognition and linking methods and sequential linguistic pipeline, where the two stages are separated, for addressing the challenge task. We describe the evaluation process and discuss the performance of the different approaches that have entered the 2016 NEEL Challenge. We also release, with this paper, the corpus consisting of manually annotated tweets.

Keywords

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

1. INTRODUCTION

Tweets are short and informal 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 such as music recommendation, social bots, spam detection, emergency response. However, extracting words and linking them to public informative resources present various challenges, due, among others, to the inherent characteristics of this type of data:

i) the restricted length and context of the message;

ii) the noisy lexical nature of the text, where terminology differs between users when referring to the same thing, and non-standard abbreviations are used.

A popular approach for making sense of tweets 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 tweets has, however, proven to be a challenging task. This was the focus of the Concept Extraction (CE) Challenge in 2013 [5]. 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 years, from the 2014 [4], 2015 [22] until the 2016 current edition, which represents a consolidation of the previous years' challenge in terms of tasks and in setting up an open and competitive environment that would encourage participants to deliver novel or improved approaches for recognizing and linking entities from tweets to either a reference Knowledge Base entry or NIL when such a reference does not exist. To encourage competition, we solicited sponsorship for the winning submission, an award of €750. This was provided by the FREME project,[1] an European H2020[2] project that aims to develop an open framework of services for multilingual and semantic enrichment of digital content ready to be used by digital content managers. These technologies are capable to process (harvest and analyse) content, capture datasets, and add value throughout content and data value chains across sectors, countries, and languages. This generous sponsorship is testament of 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 2016 NEEL Challenge, detailing its rationale and research challenges, the collaborative annotation of the corpus, 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 tweets for their own objectives.

[1] http://www.freme-project.eu
[2] https://ec.europa.eu/programmes/horizon2020

2. TASK DEFINITION AND EVALUATION

In this section, we describe the goal and tasks of the challenge, and the annotation guidelines we followed to generate the NEEL 2016 corpus of Microposts.

2.1 Task and Research Challenges

The 2016 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 2015-04 dataset representing the same real world entity, or NIL for cases where no such entry exists;
iii) clustering of all mentions linked to NIL. Thus, the same entity, which does not have a corresponding entry in DBpedia, will be referenced with the same NIL identifier.

In the remainder of this paper, we refer to the term appearing in a text as either an *entity mention* or simply a *mention*, while we refer to its DBpedia referent as the *entity*. Consequently, the operation of entity mention 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, an entity mention 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 2016 challenge, as opposed to the previous edition, not all noun phrases are considered as entity mentions. E.g.,

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 instance of an Organization, 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.,

Tweet:
head of sharm el sheikh hospital is DENYING

"head of sharm el sheikh hospital" refers to a Person; 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:

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 typed based on the context

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

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, while the second [`Apple`] mention refers to an Organisation.

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.[5] 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 allowed to submit up to three runs of their system on the test data. The evaluation was conducted using three different metrics:

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

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) [15] 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.

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

The TAC KBP 2014 scorer[6] was used to perform the evaluation.

[5] Note that in this case "Chicago" is not a qualifier, but rather, part of the entity mention.

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

3. CORPUS CREATION AND ANNOTATION

In this section, we describe the challenge dataset and the annotation process. Since the challenge task was to automatically recognize, 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 2016 NEEL Task Definition (Section 2) builds upon the previous 2014 and 2015 challenges. This consolidated both the task and extended the previously published corpus, with the only difference being the DBpedia version.[7]

In particular, the training corresponds the entire corpus of the NEEL 2015 challenge (as-is) and consists of tweets published in 2011, 2013, 2014, and 2015. Tweets from 2011 and 2013 were extracted from a collection of over 18 million tweets provided by the Redites project.[8] 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 2014 and 2015 covering both event (such as the UCI Cyclo-cross World Cup, Star Wars The Force Awakens Premiere) and non-event tweets. The development and test datasets for the 2016 challenge were created by adding tweets collected in December 2015 around the US primary elections and the Star Wars The Force Awakens Premiere.

Table 1: General statistics of the 2016 NEEL corpus. Dev refers to the Development set. *tweets* refers to the number of tweets in the set; *words* refers to the unique number of words, thus without repetition; *tokens* refers to the total number of words; *tokens/tweet* represents the average number of tokens per tweet, *entities* refers to the unique number of named entities including NILs; *NILs* refers to the number of entities not yet available in the knowledge base; *total entities* corresponds to the number of entities with repetition in the set; *entities/tweet* refers to the average of entities per tweet; *NILs/tweet* corresponds to the average of NILs per tweet. * only 300 tweets have been randomly selected to be annotated. + figures refer to the 300 tweets sampled.

	Training	Dev	Test
tweets	6,025	100	3,164
words	26,247	841	13,728
tokens	67,393	1,406	45,164
tokens/tweet	16.61	14.06	14.27
entities	3,833	174	430*
NILs	2,291	85	284*
total entities	8,665	338	1,022*
entities/tweet	1.43	3.38	3.412+
NILs/tweet	0.38	0.85	0.95+

3.1 Corpus Description

The corpus consists of three main datasets: Training (64.86%), Development (1.08%) – which enabled participants to tune their systems – and Test (34.06%). The statistics describing the data are provided in Table 1.[9] The Training set comprises of 6,025 tweets, with 67,393 tokens and 8,664 total entities. This dataset corresponds to the entire corpus of the 2015 NEEL challenge[10] (Training + Dev + Test sets).

The Development dataset consists of 100 tweets, with 1,406 tokens and 338 named entities. The Test set consists of 3,164 tweets and contains 45,164 tokens. From the Test set, we have selected a random portion of 300 tweets, which we manually annotated, totalling 1,022 total entities. We observe a similar distribution of entities per tweet for Dev and Test sets, while a different distribution for the Training set. This is the same trend for the distribution of NILs per tweet.

Summary statistics of the entity types are provided in Table 2. Across the 3 datasets, the most frequent types are Person, Organization and 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-entities are distributed fairly evenly across the three datasets. The distributional differences between the entity types in the three sets can be clearly seen. This makes the 2016 NEEL task challenging, particularly when tackled with supervised learning approaches.

Table 2: Entity type statistics for the three data sets. Dev refers to the Development set.

Type	Training	Dev	Test
Character	0.73%	5.62%	5.58%
Event	5.56%	2.07%	2.35%
Location	21.56%	5.03%	4.21%
Organization	18.94%	9.76%	15.46%
Person	32.84%	35.50%	32.97%
Product	13.84%	37.87%	34.74%
Thing	6.58%	4.14%	4.79%

3.2 Generating the Gold Standard

From the newly collected tweets for the 2016 challenge, a stratified sample that consisted of both the US primary elections and the Star Wars premiere were selected. The Development set consists of 100 tweets; the Test set comprises an initial set of 3,164 tweets, from which a sample of 300 tweets was selected to be manually annotated, though participants were asked to process the entire set of 3,164 tweets: this to enforce fairness in the evaluation procedure. The annotation environment in GATE[11] with the ontology plugin was used to mark the entity and event mentions and provide the entity types and links.

Two annotators annotated all tweets, such that difficult cases could be identified and resolved. The inter-annotator agreement (IAA) was computed using the annotation diff tool in GATE. As the annotators are not only classifying predefined mentions but can also

[7] For this 2016 year challenge is DBpedia 2015-04.

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

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

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

[11] http://gate.ac.uk

Table 3: Inter-Annotator Agreement on the Gold Standard Development and Test datasets

	Precision	Recall	F-measure
Dev.	85.84	88.72	87.26
Test	94.77	95.82	95.29

define different mentions, traditional IAA measures such as Cohen's Kappa are less suited to this task. Therefore, we measure the IAA in terms of precision, recall and F-measure[8]. The scores are presented in Table 3.

Difficult cases and disagreements were discussed among the annotators and after which the annotations were corrected. Any final disagreements were resolved manually by the most experienced annotator.

3.3 Corpus Quality

As natural language is highly expressive, it is difficult to create a well-balanced corpus. After creating the NEEL 2016 gold standard corpus, we started analyzing the dataset according to different characteristics such as confusability and readability to assess the quality and coverage of the created dataset. For this, we reuse measures and scripts from [26]. The analyses indicate that more attention needs to be paid to the language variation in the corpus, but with the history of the NEEL Challenges and its building upon previous datasets, this can be overcome.

3.4 Confusability

We define the true confusability of a surface form s as the number of meanings that this surface form can have.[12] Because new organisations, people and places are named every day, there is no exhaustive collection of all named entities in the world. Therefore, the true confusability of a surface form is unknown, but we can estimate the confusability of a surface form through the function $A(s) : S \Rightarrow \mathbb{N}$ that maps a surface form to an estimate of the size of its candidate mapping, such that $A(s) = |C(s)|$.

The confusability of a location name offers only a rough *a priori* estimate of the difficulty in linking that surface form. Observing the annotated occurrences of this surface form in a text collection allows us to make more informed estimates. We show the average number of meanings denoted by a surface form, indicating the confusability, as well as complementary statistical measures on the datasets in Table 4. In this table, we observe that most datasets have a low number of average meanings per surface form, but there is a fair amount of variation, i.e. number of surface forms that can refer to a meaning.

3.5 Dominance

We define the true dominance of a resource r_i for a given surface form s_i to be a measure of how commonly r_i is meant with regard to other possible meanings when s_i is used in a sentence. Let the dominance estimate $D(r_i, s_i)$ be the relative frequency with which the resource r_i appears in Wikipedia links where s_i appears as the anchor text. Formally:

$$D(r_i, s_i) = \frac{|WikiLinks(s_i, r_i)|}{\forall_{r \in R}|WikiLinks(s_i, r)|}$$

[12]As surface form, we refer to the lexical value of the mention.

Table 4: Confusability stats for analysed datasets. Average stands for average number of meanings per surface form, Min. and Max. stand for the minimum and maximum number of meanings per surface form found in the corpus respectively, and σ denotes the standard deviation.

Corpus	Average	Min.	Max.	σ
NEEL 2014	1.02	1	3	0.16
NEEL 2015	1.05	1	4	0.25
NEEL 2016	1.04	1	3	0.22

The dominance statistics for the analysed datasets are presented in Table 5. The dominance scores for all corpora are quite high and the standard deviation is low, meaning that in the vast majority of cases, a single resource is associated with a certain surface form in the annotations, creating a low variance for an automatic disambiguation system.

Table 5: Dominance stats for analysed datasets.

Corpus	Dominance	Max	Min	σ
NEEL 2014	0.99	47	1	0.06
NEEL 2015	0.98	88	1	0.09
NEEL 2016	0.98	88	1	0.08

3.6 Readability

To gain an understanding of the difficulty of a text, several readability measures have been developed. In this subsection, we describe the most common measures. The scores for each on the NEEL corpora are presented in Table 6.

Flesch-Kincaid [14] Initially the Flesch-Kincaid measure was developed by the US Navy to estimate the difficulty of technical manuals. It is currently often used for official documents such as those in the law and insurance domain. Its score corresponds to a US school grade level and is computed as:

$$11.8 * \frac{syllables}{words} + 0.39 * \frac{words}{sentences} - 15.59 \quad (2)$$

Automated Readability Index (ARI) [23] The ARI index was also developed by the US military, and contrary to the Flesch-Kincaid test it compares characters to gauge the word length instead of syllables. The obtained scores correspond to US school grade levels. Decimal scores are rounded up. It is computed as:

$$4.71 * \frac{characters}{words} + 0.5 * \frac{words}{sentences} - 21.43 \quad (3)$$

Coleman-Liau [7] Similar to ARI, Coleman-Liau uses characters instead of syllables. It also roughly corresponds to US school grade levels. It is computed as:

$$5.88 * \frac{characters}{words} - 29.5 * \frac{sentences}{words} - 15.8 \quad (4)$$

Flesch Reading Ease [10] The Flesch Reading Ease score was developed by Rudolf Flesch in 1979. In this index, the scores lie between 0.00 and 100.0 where a higher score indicates an easier text to read. The formula is as follows:

$$206.835 - 84.6 * \frac{syllables}{words} - 1.015 * \frac{words}{sentences} \quad (5)$$

Table 6: Readability scores for analysed datasets.

Corpus	Flesch Kincaid	ARI	Coleman-Liau	Flesch Index	Fog Index	LIX	SMOG
NEEL 2014	5.9	6.4	7.6	79.8/100	8.9	32.7	8.6
NEEL 2015	6.0	6.4	7.5	79.7/100	9.0	32.6	8.7
NEEL 2016	6.0	6.7	8.6	76.4/100	8.9	33.7	8.9

Fog Index [13] The FOG index was created by businessman Robert Gunning and discerns between the proportion of sentences with 'easy' and 'difficult' words. This difficulty is defined by the number of syllables a word has, although one could argue that long frequent words are less difficult than short infrequent words. It is computed using the following formula and its score also corresponds to the number of years of education deemed necessary to understand a text.

$$0.4 * \frac{words}{sentences} + 100 * \frac{words >= 3syllables}{words} \quad (6)$$

LIX [1] LIX was developed in Sweden. Rather than relying on the number of syllables or absolute character counts to distinguish long words, it computes the proportion of words that are over 6 characters and it is one of the few readability measures developed for languages other than English. The score can range between 20 and 60, with a higher score indicating that a text is more difficult. It is computed as:

$$\frac{words}{sentences} + 100 * \frac{words >= 6characters}{words} \quad (7)$$

SMOG grading [16] The SMOG reading formula was developed as a fix to the Fog index. It is computed as:

$$\sqrt{\frac{words >= 3syllables}{sentences} * 30} + 3 \quad (8)$$

Generally, the readability scores would indicate that tweets are fairly easy to understand, as grade levels around 6 are deemed suitable for 10-11 year-olds. However, applying these readability measures to tweets uncovers their main weakness, namely that while tweets do contain shorter words and sentences in general, they also contain many abbreviations and cultural terms. None of the readability measures investigated is equipped to deal with this.

3.7 Summary

In this section, we have analysed the corpora in terms of their variance in named entities and readability.

As the datasets are built on top of each other, they show a fair amount of overlap in entities between each other. This is not a problem, if there is enough variation among the entities, but the confusability and dominance statistics show that there are very few entities in our datasets with many different referents ('John Smiths') and if such an entity is present, often only one of its referents is meant. To remedy this, future entity linking corpora should take care to balance the entity distribution and include more variety.

As for the readability of the different datasets, the readability measures indicate that tweets are generally not very difficult in terms of word and sentence length, but the abbreviations and slang present in tweets proves them to be more difficult to interpret for readers outside the target community. To the best of our knowledge, there is no readability metric that takes this into account.

Next to the thorough analysis of the corpora, we also make use of a baseline system, namely, the ADEL entity linking framework [18]. ADEL represents a good state-of-the art system which has been used to discover potential annotation inconsistencies such as *i)* missing extracted entities, *ii)* wrongly typed entities and *iii)* wrongly linked entities that have been corrected for the benefits of all participants.

4. PARTICIPANT OVERVIEW

The challenge attracted a lot of interest from research groups spread around the world. Thirty-seven groups expressed their intent to participate to the challenge and acquired the corpus. Table 7 lists the teams that participated in the final evaluation of the challenge task. In this section, we first present the baseline system used to improve the quality of the annotations (Section 4.1) and as a means to compare the participant systems' results. We then present the approaches followed by each participant system (Section 4.2).

4.1 Baseline System

We used ADEL [19], which combines linguistic and graph-based algorithms to detect entity mentions in text and to link them to referents in a knowledge base, as a baseline system in order to assess the quality of the dev and test sets. This hybrid annotator consists of three main stages: *i)* Mention Extraction, *ii)* Resolution and Classification, and *iii)* Candidate Selection.

i) This stage detects mentions in text that are likely to denote entities. It is composed of extractors module that make use of dictionaries, Part-of-Speech (POS) tagging nouns, and Named Entity Recognition (NER) classifying entities. In details, we use the Stanford NLP POS-Tagger [25] with the *english-bidirectional-distsim* model and the Stanford NER Tagger [9] using the *NERClassifierCombiner* functionality to combine multiple CRF models together.

ii) The parallel strategy used in the mention extraction may generate overlaps. We use an *overlap resolution* module that takes the output of each component of the extractors module and decides on a single output with no overlaps. The type of the mention is then assigned according to the type assigned on the match or partial match by the entity recognizer.

iii) This stage aims to propose candidate links that are later on ranked, the first ranked link being the most suitable resource to disambiguate the entity. We perform a lookup for an entity mention in an index built on top of both DBpedia2015-04[13] and a dump of the Wikipedia articles from February 2015[14] to generate potential candidates for an entity mention. The ranking function *r(l)* combines: *i)* the Levenshtein distance L between the entity mention m and a knowledge base label (e.g. Wikipedia title), *ii)* the maximum Levenshtein distance between the mention m and a label (title) of every Wikipedia redirect page from a set R, *iii)* the maximum distance between the mention m and every label (title) in the set

[13] http://wiki.dbpedia.org/services-resources/datasets/datasets2015-04

[14] https://dumps.wikimedia.org/enwiki

Table 7: Accepted submissions with team affiliations together with the number of runs used. * indicates the baseline system.

Reference	Team's affiliation	Team Name	Authors	No. of runs
[27]	Hasso-Plattner-Institute Potsdam, Germany	kea	Waitelonis *et al.*	1
[24]	Insight Centre for Data Analytics, National University of Ireland, Ireland	insight-centre @ nuig	Torres *et al.*	1
[12]	MIT Lincoln Laboratory, US	mit lincoln laboratory	Greenfield *et al.*	2
[11]	Jadavpur University, India	ju team	Gosh *et al.*	3
[3]	University of Milano-Bicocca, Italy	unimib	Caliano *et al.*	2
[19]	EURECOM, France and ISMB, Italy	adel*	Plu *et al.*	1

of Wikipedia disambiguation pages D, *iv)* the PageRank [17] PR value for every entity candidate l.[15] Entities with no detected entry in a knowledge base are assigned *NIL* links. In cases when two or more candidate links attached to a single entity mention share the same maximum ranking score, we still rely on taking the first candidate in the ordered ranking list as the most probable candidate.

We then used ADEL to detect annotation inconsistencies through an analysis mode which, for a given test set and its gold standard, provides the differences between the gold standard annotations and the system annotations. We use this debug mode to highlight the differences between the gold standard annotations and the results provided by ADEL and we manually review those differences.

We performed this operation on both the Dev and Test sets. We found 24 annotations issues in the dev set but 0 in the test set. They belong roughly to four categories: the multiple re-tweets of some tweets were not all annotated; some entities were not extracted; some surface forms were controversial (e.g. the mention "Star Wars Episode V: The Empire Strikes Back" was wrongly split into the two entities "Star Wars" and "The Empire Strikes Back"); some entities were wrongly linked (e.g. @carrieffisher has been linked to NIL instead of `dbp:Carrie_Fisher`). These issues have been then submitted to two experts, who have agreed on the errors and issued a corrected version of the dev set.

4.2 Participant Systems

Waitelonis *et al.* [27] propose a joint Mention Extraction and Candidate Selection, where ngrams of the text are mapped to DBpedia entities. A preprocessing stage cleans and normalizes the initial tweets; scoring measures, weighting graph distance measurements, connected component analysis, centrality of the entities and density observations are used to resolve the selection of entities in case of ambiguity. The candidate selection is sorted according to the confidence score, which is used as means to understand whether the entity actually describes the mention. In case the confidence score is lower than an empirically threshold, the mention is annotated with a NIL.

Torres *et al.* [24] implement a linguistic pipeline where the Candidate Selection is performed by looking up entities according to the exact lexical value of the mentions with DBpedia titles, redirect pages, and disambiguation pages. A crucial part of the approach is the preprocessing that consists of normalizing the input text and making it similar to formal language text. The entity extraction and typing is performed by the GATE NLP Framework [8]. The final disambiguation process takes as input a list of named entities, each containing a list of candidate DBpedia resources after the linking stage, and applies entity relatedness reasonings. The output is the selection of the best candidate resource for each input named entity. For the entities linked to the mentions, but without a corresponing referent in the knowledge base, authors apply a incremental and hierarchical clustering approach: they iterate over each NIL-linked entity and aggregate them into clusters one by one. The first element is assigned into an initial cluster, then the next item is compared to the previous ones using the Monge-Elkan similarity measure [6].

Greenfield *et al.* [12] propose a joint graph-based and linguistic approach, without performing any tweet normalization. DBpedia is used as a dictionary of entities. The authors mapped the DBpedia Ontology to the NEEL challenge taxonomy (Appendix A) of entity classes; this mapping resulted to be not thorough for the Dev set ranging from the 100% of the Person-type entities to only 11% of Character. Authors apply then a parallel candidate name generation. The linking is turned as a binary classification task. An extensive feature set is used, where relevant features are: COMMONNESS, IDFanchor, TEN, TCN, TFsentence, TFparagraph, and REDIRECT. The entity type is assigned via NER, based on CRF. The final clustering is performed using the normalized Damerau-Levenshtein, holding better performance than Brown clustering [2].

Gosh *et al.* [11] implement a sequential linguistic pipeline composed of Preprocessing, Named Entity Recognition (NER), Linking (NEL), and, finally, clustering. For the preprocessing stage, they enrich the set of mentions, using the stratified bag of entities (grouped by DBpedia types) from the training set to gather additional mentions. The recognition is performed using both Stanford NER and ARK Twitter Part-of-Speech Tagger to detect proper nouns. The classification of the extracted nouns is performed by a random forest using a rich feature vector. Among the used features, we mention: length of the mention, capitalization of the mention, mention if it contains mixed cases, mention if it contains digits, if period is in the phrase (mention), frequency of the POS mention, if mention is the Person list, or in the Event list, or in the Character list (the lists are built from Wikipedia). The NEL is performed querying the off-the-shelf Babelfy, annotating one tweet by time and matching with the entities defined in the previous step. The NIL is achieved via a clustering of the unlinked entities, performed via an exact match of the mentions.

Caliano *et al.* [3] propose a sequential approach composed of entity identification, candidate selection and ranking, entity linking and

[15]The PageRank scores for every DBpedia resource originate from [20].

typing, and a final stage of entity mention re-scoping. For the entity identification, they first remove special characters such as #, @, and then used T-NER [21] off-the-shelf. Then a learning to rank strategy is applied for the candidate selection, where a linear regression weights the lexical similarity of the mention with the Wikipedia title and the contextual similarity of the text surrounding the mention and the DBpedia abstract. The candidate resource with the highest candidate score is selected to be assigned as final entity. Typing is performed inheriting the DBpedia class, previously mapped to the taxonomy used in the NEEL challenge. A final post-processing stage is implemented to fixing mention boundary problems.

5. CHALLENGE RANKING

Table 8 provides the 2016 NEEL challenge ranking. 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 *kea* [27] outperformed all other participants, with an overall performance score of 0.5486 and a delta difference of 16.58% with respect to the second ranked approach. The top-ranked system performed, however, lower than the chosen baseline (ADEL). All ranked systems underline current and ongoing research and industrial path in pushing toward a hybrid graph-based and linguistic approach, where the NIL detection is the direct output of the disambiguation stage when the confidence score does not satisfy a minimal threshold.

Table 8: Final Ranking of the 2016 NEEL challenge. * is used as baseline thus not ranked.

Rank	Reference	Team Name	r_S
1	**[27]**	**kea**	**0.5486**
2	[24]	insight-centre @ nuig	0.3828
3	[12]	mit lincoln laboratory	0.3609
4	[11]	ju team	0.3548
5	[3]	unimib	0.3353
*	[19]	adel	0.6198

Table 9 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.

Table 9: Breakdown mention_ceaf figures per participant. * is used as baseline thus not ranked.

Rank	Reference	Team Name	F_1
1	[27]	kea	0.641
2	[24]	insight-centre @ nuig	0.621
3	[11]	ju team	0.467
4	[12]	mit lincoln laboratory	0.366
5	[3]	unimib	0.203
*	[19]	adel	0.69

Table 10 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 11 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 10: Breakdown strong_typed_mention_match figures per participant. * is used as baseline thus not ranked.

Rank	Reference	Team Name	F_1
1	[27]	kea	0.473
2	[12]	mit lincoln laboratory	0.319
3	[11]	ju team	0.312
4	[3]	unimib	0.267
5	[24]	insight-centre @ nuig	0.246
*	[19]	adel	0.61

Table 11: Breakdown strong_link_match figures per participant. * is used as baseline thus not ranked.

Rank	Reference	Team Name	F_1
1	[27]	kea	0.501
2	[12]	mit lincoln laboratory	0.396
3	[11]	ju team	0.248
4	[24]	insight-centre @ nuig	0.202
5	[3]	unimib	0.162
*	[19]	adel	0.536

6. CONCLUSIONS

The 2016 NEEL challenge aims to foster the development of novel approaches for mining information from tweets and linking it to external knowledge. The motivation for organizing this challenge is the strong interest of the research and commercial communities in developing systems able to fit the challenging context of mining semantics from tweets, in particular the challenging tasks of 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.

As in 2015, we used the evaluation metrics proposed in TAC KBP 2015 tasks to account for *mention_ceaf*, *strong_link_match*, and *strong_typed_mention_match*. Carrying out evaluation in this way provides a robust and standardized approach for ranking participants' entries.

As a result of the 2016 NEEL challenge, we have generated a manually annotated corpus, which extends the 2014 and 2015 challenges with the annotations 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 tweets.

The gold standard[16] is released with the CC BY 4.0 license.[17] We hope that through our release of data and resources, we will promote research on entity recognition and disambiguation, especially with regard to tweets.

This year challenge has underlined and consolidated the awareness on the use of joint graph-based and linguistic approaches to cope with the challenge task. The extensive use of a well-defined encyclopedic graph gives a better understanding of the tweet context, thus helping to better define the final knowledge base referent resource. Performance results still show the complexity of the task,

[16] PLEASE-ABA-SHA-PUT-THE-LINK

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

that has been increased compared with previous year due to the larger set of data being processed by the participants. We can conservatively claim that NEEL over tweets is still an open research challenge.

As in 2015, also in 2016, we 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

The authors would like to thank Bianca Pereira for her support and engagement in shaping the current and future research activities concerning the NEEL challenge. Special thanks to the FREME project (GA No. 644771) who generously sponsored the prize for the winning submission. We thank also the participants who helped to improve the corpus. This work was supported primarily by the FREME project (GA no. 644771), NewsReader (GA no. ICT-316404), and the CLARIAH-CORE project financed by NWO (`http://www.clariah.nl`).

8. REFERENCES

[1] C.-H. Björnsson. *Läsbarhet*. Liber, 1968.

[2] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai. Class-based N-gram Models of Natural Language. *Comput. Linguist.*, 18(4):467–479, 1992.

[3] D. Caliano, E. Fersini, P. Manchanda, M. Palmonari, and E. Messina. UniMiB: Entity Linking in Tweets using Jaro-Winkler Distance, Popularity and Coherence. In *6th International Workshop on Making Sense of Microposts (#Microposts)*, 2016.

[4] 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 (#Microposts)*, 2014.

[5] A. E. Cano Basave, A. Varga, M. Rowe, M. Stankovic, and A.-S. Dadzie. Making Sense of Microposts (#MSM2013) Concept Extraction Challenge. In *3rd Workshop on Making Sense of Microposts (#MSM)*, 2013.

[6] W. W. Cohen, P. Ravikumar, and S. E. Fienberg. A Comparison of String Metrics for Matching Names and Records. In *KDD Workshop on Data Cleaning and Object Consolidation*, 2003.

[7] M. Coleman and T. L. Liau. A computer readability formula designed for machine scoring. *Journal of Psychology*, 60:283–284, 1975.

[8] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, W. Peters, L. Derczynski, and et al. Developing Language Processing Components with GATE Version 8 (a User Guide). Technical report, The University of Sheffield, Department of Computer Science, 2014.

[9] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *43rd Annual Meeting on Association for Computational Linguistics*, 2005.

[10] R. Flesch. *How to Write Plain English: A Book for Lawyers and Consumers*. 1979, Harper & Row.

[11] S. Ghosh, P. Maitra, and D. Das. Feature Based Approach to Named Entity Recognition and Linking for Tweets. In *6th International Workshop on Making Sense of Microposts (#Microposts)*, 2016.

[12] K. Greenfield, R. Caceres, M. Coury, K. Geyer, Y. Gwon, J. Matterer, A. Mensch, C. Sahin, and O. Simek. A Reverse Approach to Named Entity Extraction and Linking in Microposts. In *6th International Workshop on Making Sense of Microposts (#Microposts)*, 2016.

[13] R. Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.

[14] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training, U. S. Naval Air Station, Memphis, TN, 1975.

[15] X. Luo. On Coreference Resolution Performance Metrics. In *Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 2005.

[16] G. H. McLaughlin. Smog grading – a new readability formula. *Journal of Reading*, 12(8):639–646, 1969.

[17] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the web. Technical report, Stanford InfoLab, 1999.

[18] J. Plu, G. Rizzo, and R. Troncy. Revealing Entities from Textual Documents Using a Hybrid Approach. In *3rd International Workshop on NLP & DBpedia*, Bethlehem, Pennsylvania ,USA, 2015.

[19] J. Plu, G. Rizzo, and R. Troncy. Enhancing Entity Linking by Combining NER Models. In *13th Extended Semantic Web Conference (ESWC), Challenges Track*, 2016.

[20] D. Reddy, M. Knuth, and H. Sack. DBpedia GraphMeasures. dataset, 2014.

[21] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named Entity Recognition in Tweets: An Experimental Study. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2011.

[22] G. Rizzo, E. C. Cano Amparo, B. Pereira, and A. Varga. Making sense of Microposts (#Microposts) Named Entity rEcognition & Linking Challenge. In *5th International Workshop on Making Sense of Microposts (#Microposts)*, 2015.

[23] R. J. Senter and E. A. Smith. Automated readability index. Technical report, Wright-Patterson Air Force Base, 1965.

[24] P. Torres-Tramon, H. Hromic, B. Walsh, B. Heravi, and C. Hayes. Kanopy4Tweets: Entity Extraction and Linking for Twitter. In *6th International Workshop on Making Sense of Microposts (#Microposts)*, 2016.

[25] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, 2003.

[26] M. van Erp, P. Mendes, H. Paulheim, F. Ilievski, J. Plu, G. Rizzo, and J. Waitelonis. Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In *10th International Conference on Language Resources and Evaluation (LREC)*, 2016.

[27] J. Waitelonis and H. Sack. Named Entity Linking in #Tweets with KEA. In *6th International Workshop on Making Sense of Microposts (#Microposts)*, 2016.

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]
```