Undefined 1 (2009) 1–5
IOS Press



# Lessons Learnt from the Named Entity rEcognition and Linking (NEEL) Challenge Series

*Emerging Trends in Mining Semantics from Tweets*

Giuseppe Rizzo [a,*], Bianca Pereira [b] and Andrea Varga [c] and Marieke van Erp [d] and Amparo Elizabeth Cano Basave [e]

[a] *ISMB, Turin, Italy. E-mail: giuseppe.rizzo@ismb.it*
[b] *The Insight Centre for Data Analytics, Galway, Ireland. E-mail: bianca.pereira@insight-centre.org*
[c] *The Content Group, Godalming, UK. E-mail: varga.andy@gmail.com*
[d] *Vrije Universiteit Amsterdam, Netherlands. E-mail: marieke.van.erp@vu.nl*
[e] *Cube Global, United Kingdom. E-mail: ampaeli@gmail.com*

**Abstract.** The large number of tweets generated daily is providing policy makers with means to obtain insights into recent events around the globe in near real-time. The main barrier for extracting such insights is the impossibility of manual inspection of a diverse and dynamic amount of information. This problem has attracted the attention of industry and research communities, resulting in algorithms for the automatic extraction of semantics in tweets and linking them to machine readable resources. While a tweet is shallowly comparable to any other textual content, it hides a complex and challenging structure that requires domain-specific computational approaches for mining semantics from it. The NEEL challenge series, established in 2013, has contributed to the collection of emerging trends in the field and definition of standardised benchmark corpora for entity recognition and linking in tweets, ensuring high quality labelled data that facilitates comparisons between different approaches. This article reports the findings and lessons learnt through an analysis of specific characteristics of the created corpora, limitations, lessons learnt from the different participants and pointers for furthering the field of entity recognition and linking in tweets.

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

## 1. Introduction

Tweets have been proven to be useful in different applications and contexts such as music recommendation, spam detection, emergency response, market analysis, and decision making. The limited number of tokens in a tweet however implies a lack of sufficient contextual information necessary for understanding its content. A commonly used approach is to extract *named entities*, which are information units such as the names of a Person or an Organisation, a Location, a Brand, a Product, a numerical expression including Time, Date, Money and Percent found in a sentence [38], and enrich the content of the tweet with such information. In the context of the NEEL challenge series, we extended this definition of *named entity* as being a phrase representing the name, excluding the preceding definite article (i.e. "the") and any other pre-posed (e.g. "Dr","Mr") or post-posed modifiers, that belongs to a class of the NEEL Taxonomy (ref. Appendix) and are linked to a DBpedia resource. The semantically enriched tweet have been shown to help

*Corresponding author. E-mail: giuseppe.rizzo@ismb.it

addressing complex information seeking behaviour in social media, such as semantic search [85], deriving user interests [87], and disaster detection [86].

The automated identification, classification and linking of named entities has proven to be challenging due to, among other things, the inherent characteristics tweets: *i)* the restricted length and *ii)* the noisy lexical nature, i.e. terminology differs between users when referring to the same thing, and non-standard abbreviations are common. Numerous initiatives have contributed to the progress in the field broadly covering different types of textual content (and thus going beyond the boundaries of tweets). For example TAC-KBP [45] has established a yearly challenge in the field covering newswire, websites, discussion forum posts, ERD [49] with search queries content, and SemEval [50] with technical manuals and reports.

The NEEL challenge series, established first in 2013 and since then running yearly, has captured a community need for making sense from tweets through a wealth of high quality annotated corpora and to monitor the emerging trends in the field. The first edition of the challenge named Concept Extraction (CE) Challenge [1] focused on entity identification and classification. A step further into this task 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 [2]. These two research avenues, which add to the intrinsic complexity of the tasks proposed in 2013 and 2014, prompted the Named Entity rEcognition and Linking (NEEL) Challenge in 2015 [3]. In 2015, the role of the named entity type in the grounding process was investigated, as well as the identification of named entities that cannot be grounded because they do not have a knowledge base referent (defined as NIL). The English DBpedia 2014 dataset was the designated referent knowledge base for the 2015 NEEL challenge, and the evaluation was performed through live querying the Web APIs participants prepared, in an automatic fashion to measure the computing time. The 2016 edition [4] consolidated the 2015 edition, using the English DBpedia 2015-04 version as referent knowledge base. This edition proposed an offline evaluation where the computing time was not taken into account in the final evaluation.

The four challenges have published four incremental manually labeled benchmark corpora. The creation of the corpora followed rigid designations and protocols, to grant high quality labeled data that can be used as seeds for reasoning and supervised approaches. Despite these protocols, the corpora have strengths and weaknesses that we have discovered over the years and they are discussed in this article.

The purpose of each challenge was to set up an open and competitive environment that would encourage participants to deliver novel approaches or improve on existing ones for recognising and linking entities from tweets to either a referent knowledge base entry or NIL where such an entry does not exist. From the first (in 2013) to the 2016 NEEL challenge, thirty research teams have submitted at least one entry to the competitions proposing state-of-the-art approaches. More than three hundred teams have explicitly acquired the corpora in the four years, underlining the importance of the challenges in the field.[1] The NEEL challenges have also experienced a strong involvement of the industry as both participants and funding agencies. For example, in 2013 and 2015 the best performing systems were proposed by industrial participants. The prizes were sponsored by industry (ebay[2] in 2013 and SpazioDati[3] in 2015) and research projects (LinkedTV[4] in 2014, and FREME[5] in 2016). The NEEL challenge also triggered the interest of localised challenges such as the NEEL-IT, the NEEL challenge for tweets written in Italian [88] that brings the multilinguality aspect in the NEEL contest.

This paper reports on the findings and lessons learnt from the last four years of NEEL challenges, analysing the corpora in detail, highlighting their limitations, and providing guidance to implement top performing approaches in the field from the different participants. 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. The remainder of this paper is structured as follows: in Section 2 we introduce a comparison with recent shared tasks in entity recognition and linking and underline the reason that has prompted the need to establish the NEEL challenge series. Next, in Section 3, the decisions regarding different versions of the NEEL challenge are introduced and the initiative is compared against the other shared tasks. We then detail the steps followed in generating the four different corpora in Section 4, followed by a quantitative

[1] This number does not account for the teams who experimented with the corpora out of the challenges' timeline.

[2] http://www.ebay.com

[3] http://www.spaziodati.eu

[4] http://www.linkedtv.eu

[5] http://freme-project.eu/

and qualitative analysis of the corpora in Section 5. We then list the different approaches presented and narrow down the emerging trends in Section 6, grounding the trends according to the evaluation strategies presented in Section 7. Section 8 reports the participants' results and provides an error analysis. We conclude and list our future activities in Section 9.

## 2. Entity Linking Background

The first research challenge to identify the importance of the recognition of entities in textual documents was held in 1997 during the $7^{th}$ Message Understanding Conference (MUC-7) [44]. In this challenge, the term *named entity* was used for the first time to represent terms in text that refer to instances of classes such as Person, Location, and Organisation. Since then, named entities have become a key aspect in different research domains, such as Information Extraction, Computational Linguistics, Machine Learning, Semantic Web.

Recognising an entity in a textual document was the first big challenge, but after overcoming this obstacle, the research community moved into a second and challenging task: disambiguating entities. This problem appears when a mention in text may refer to more than one entity. For instance, the mention *Paul* appearing in text may refer to the singer *Paul McCartney*, to the actor *Paul Walker*, or to any of the millions of people called *Paul* around the world. In the same manner, *Copacabana* can be a mention of the beach in *Rio de Janeiro, Brazil*, or the beach in *Dubrovnik, Croatia*. The problem of ambiguity is translated into the question "which is the exact entity that the mention in text refers to?". To solve this problem, recognising the mention to an entity in text is only the first step for the semantic processing of textual documents, the next one is to ground the mention to an unambiguous representation of the same entity in a knowledge base. This task became known in the research community as Entity Disambiguation.

The Entity Disambiguation task popularised after Bunescu and Pasca [40] in 2006 explored the use of an encyclopaedia as a source for entities. In particular, after [41] demonstrated the benefit of using Wikipedia,[6] a free crowd-sourced encyclopaedia, for such purpose. The reason why encyclopedic knowledge is important is that an encyclopaedia contains representation of entities in a variety of domains, and, moreover, contains a single representation for each entity along with a symbolic or textual description. Therefore from 2006 until 2009, there were two main areas of research: Entity Recognition, as a legacy of the work started during the MUC-7 challenge, and Entity Disambiguation, exploring encyclopedic knowledge bases as catalogs of entities.

In 2009, the TAC-KBP challenge [45] introduced a new problem to both the Entity Recognition and Entity Disambiguation communities. In Entity Recognition, the mention is recognised in text without information about the exact entity that is being referred by the mention. On the other hand, Entity Disambiguation focuses only on the resolution of entities that have a referent in a provided knowledge base. The TAC-KBP challenge illustrated the problem that a mention identified in text, may not have a referent entity in the knowledge base. In this case, the suggestion was to link such a mention to a NIL entity in order to indicate that it is not present in the knowledge base. This problem was referred as Named Entity Linking and it is still a hard and current research problem. Nowadays, however, the terms Entity Disambiguation and Entity Linking have been used interchangeably.

Since the TAC-KBP challenge, there has been an explosion on the number of algorithms generated to solve Entity Linking using a variety of textual documents, Knowledge Bases, and even using different definitions of entities. This variety, whilst beneficial, also extends to how approaches are evaluated, regarding metrics and gold standard datasets used. Such diversity makes it difficult to perform comparisons between various Entity Linking algorithms and creates the need for benchmark initiatives.

In this section, we first introduce the main components of the Entity Linking task and their possible variations, followed by a typical workflow used to solve the task, the expected output of each step and three strategies for evaluation of Entity Linking systems. We conclude with an overview of benchmark initiatives and their decisions regarding the use of Entity Linking components and evaluation strategies.

### 2.1. Entity Linking components

Entity Linking is defined as *the task of grounding entity mentions (i.e. words) in textual documents with knowledge base entries, in which both mention and knowledge base entry are recognised as references to*

[6] http://en.wikipedia.org

*the same entity*. If there is no knowledge base entry to ground a specific entity mention then the mention should be linked to a NIL reference instead.

In this definition, Entity Linking contains three main components: text, knowledge base, and entity. The features of each component may vary, and consequently, have an impact on the results of algorithms used to perform the task. For instance, a state-of-the-art solution based on long textual documents may have a poor performance when evaluated with short documents with little contextual information within the text. In a similar manner, a solution developed to link entities of types Person, Location, and Organisation may not be able to link entities of type Movie. Therefore, the choice for each component defines which type of solutions are being evaluated by each specific benchmark initiative.

#### *2.1.1. Textual Document*

In Entity Linking, textual documents are usually divided in two main categories: long text, and short text. Long textual documents usually contain more than 400 words, such as news articles and web sites. Short documents (such as microposts[7] or tweets) may have as few as 200 characters, or even contain a single word, as in search queries. Different types of texts have their own characteristics that may influence Entity Linking algorithms.

Long textual documents provide a series of features that can be explored for Entity Linking such as: the presence of multiple entity mentions in a single document; well-written text (expressed by the lack or low presence of misspellings); and the availability of contextual information that supports the grounding of each mention. Contextual information entails the supporting facts that help in deciding the best knowledge base entry to be linked with a given mention. For instance, let us assume the knowledge base has two candidate entries to be linked with the mention *Michael Jordan*. One of these entries refer to the *professor at University of California, Berkeley* and the other to the *basketball player*. In order to decide which is the correct entry to be linked with the text, some context needs to be provided such as: "played a game yesterday", or "won the championship", and so on. As more context is available the task becomes easier, little or no context makes the task more challenging.

[7]Microposts is the term used in the social media field to refer to tweets and social media posts in general.

Short text documents are considered more challenging than long ones because they have the exact opposite features such as: the presence of few entity mentions in a single document (due to the limited size of the text); the presence of misspellings or phonetic spelling (e.g. "I call u 2morrow" rather than "I call you tomorrow"); and the low availability of contextual information within the text. It is important to note though that even within the short text category there are still important distinctions between microposts and search queries that may impact the performance of Entity Linking algorithms.

The most striking difference concerns the intention of the user writing the document. A search engine user has a specific information need when writing a search query. She wants to be able to find exactly what she is looking for. Whereas in microposts, the goal of the user may be either statement of facts, expression of emotions, or communication of opinions.

In performing a search, it is expected that the search query will be composed by a mention to the entity of interest being searched and, sometimes, by additional contextual information. Therefore, despite the challenge of being a short text document, search queries are assumed to contain at least one mention to an entity and likely to contain additional contextual information. However, for microposts this assumption does not hold.

Microposts do not necessarily have an entity as target. For instance, a document with the content "So happy today!!!" does not explicitly cite any entity mention. Also, microposts may be used to talk about entities without providing any context within the text, as in "Adele, you rock!!". In this aspect, Entity Linking for microposts is more challenging than for search queries because it is unclear if a micropost will contain an entity and context for the linking. Furthermore, microposts are also more likely to contain misspellings and phonetic writing than search queries. If a search engine user performs a misspelling then it is very likely that she will not find the desired information. In this case, it is safe to assume that search engine users will try to avoid misspellings and phonetic writing as much as possible. On the other hand, in micropost communities, misspellings and phonetic writing are used as strategies to shorten words, thus enabling the communication of more information within a single micropost. Therefore, misspelling and phonetic writing are common features of microposts and need to be taken into consideration when performing Entity Linking.

#### 2.1.2. Knowledge Base

The second component of Entity Linking we consider is the knowledge base used. Knowledge bases differ from each other regarding the domains of knowledge they cover (e.g. domain-specific or encyclopedic knowledge), the features used to describe entries (e.g. long textual descriptions, attribute-value pairs, or relationship links between entities), and their ratio of updates.

As with textual documents, different characteristics will impact in the Entity Linking task. The domain covered by the knowledge base will influence which entity mentions will possibly have a link. If there is a mismatch between the domain of the text (e.g. biomedical text) and the domain of the knowledge base (e.g. politics) then all, or most, entity mentions found in text will not have a reference in the knowledge base. In the extreme case of complete mismatch, the Entity Linking process will be reduced to Entity Recognition. Therefore, in order to perform linking, the knowledge base should at least be partially related to the domain of the text being linked.

Furthermore, the features used to describe entities in the knowledge base influence which algorithms can make use of it. For instance, if entities are represented only through textual descriptions, a text-based algorithm needs to be used to find the best mention-entry link. If, however, knowledge base entries are only described through relationship links with other entities then a graph-based algorithm may be more suitable.

The third characteristic of a knowledge base which impacts Entity Linking is its ratio of updates. Static knowledge bases (i.e. knowledge bases that are not or infrequently updated) represent only the status of a given domain at the moment it was generated. Any entity which becomes relevant to that domain, after that point in time will not be represented in the knowledge base. Therefore, in a textual document, only mentions to entities prior to the creation of the knowledge base will have a link, all others would be linked to NIL. The faster entities change in a given domain the more likely it is for the knowledge base to become outdated. In the likelihood that there is a complete disjoint between text and knowledge base, all links from text would invariably be linked to NIL. Depending on the textual document to be linked, the ratio of updates may or may not be an important feature. Social and news media are more likely to have a faster change on their entities of interest than manufacturing reports, for instance.

#### 2.1.3. Entity

The third component of interest for Entity Linking is the definition of entity. Despite its importance for the Entity Linking task, entities are not formally defined. Instead, entities are defined either through example or through the data available. Named entities are the most common case of definition by example. Named entities were introduced in 1997 as part of the Message Understanding Conference as instances of Person, Organisation, and Geo-political types. An extension of named entities is usually performed through the inclusion of additional types such as Locations, Facilities, Movies, etc. In these cases there is no formal definition of entities, rather they are exemplars of a set of categories.

An alternative definition of entities assumes that entities are anything represented by the knowledge base. In other words, the definition of entity is given by the data available (in this case, data from the knowledge base). Whereas this definition makes the Entity Linking task easier by not requiring any refined "human-like" reasoning about types, it makes it impossible to identify NIL links. If entity is anything in the knowledge base, how could we ever possibly have, by definition, an entity which is not in the knowledge base?

The choice of entity will depend on the Entity Linking desired. If the goal is to consider links to NIL then the definition based on types is the most suitable, otherwise the definition based on the knowledge base may be used.

### 2.2. Typical Entity Linking Workflow and Evaluation Strategies

Regardless of the different Entity Linking components, most proposed systems for Entity Linking follow a workflow similar to the one presented in Figure 1. This workflow is composed of the following steps: Mention Detection, Entity Typing, Candidate Detection, Candidate Selection, and NIL Clustering. Note that, although it is usually a sequential workflow, there are approaches that create a feedback loop between different steps, or merge two or more steps into a single one.

The Mention Recognition step receives textual documents as input and recognises all terms in text that refer to entities. The goal of this step is to perform typical Named Entity Recognition. Next, the Entity Typing step detects the type of each mention previously recognised. This task is usually framed as a categorisation problem. Following, Candidate Detection receives the detected mentions and produces a list with all entries

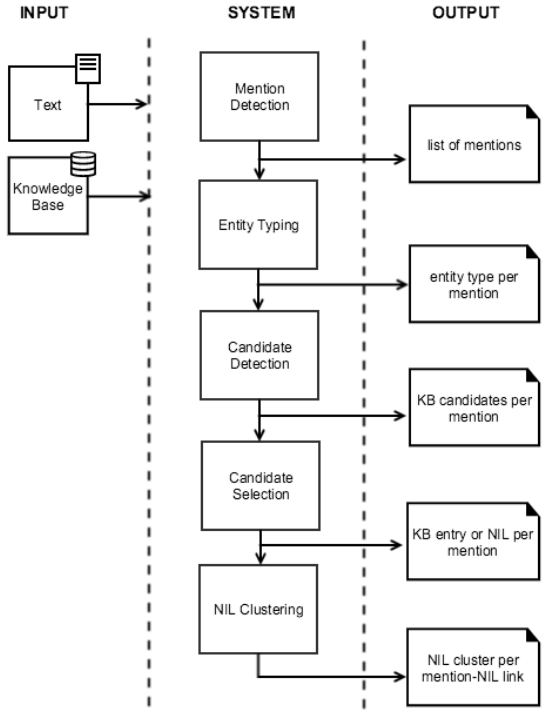


Fig. 1. Typical Entity Linking workflow with expected output of each step.

in the knowledge base that are candidates to be linked with each mention. In the Candidate Selection step, these candidate lists are processed and, by making use of available contextual information, the correct link for each mention, either an entry from the knowledge base or a NIL reference, is provided. Last, the NIL Clustering step receives a series of mentions linked to NIL as input and generates clusters of mentions referring to the same entity, i.e. each cluster contains all NIL mentions representing one, and only one, entity, and there are no two clusters representing the same entity.

The evaluation of Entity Linking systems is based on this typical workflow and can be of three types: end-to-end, step-by-step, or partial end-to-end.

An *end-to-end strategy* evaluates a system based only on the aggregated result of all its steps. It means that if one step in the workflow does not perform well and its error propagates through all subsequent steps, this type of evaluation will judge the system based only on the aggregated error. In this case, a system that performs excellent Candidate Selection but poor Mention Detection can be considered as good as a system that performs a poor Candidate Selection but an excellent Mention Detection. The end-to-end strategy is very useful for application benchmark in which the goal is to maximise the results that will be consumed by another application based on Entity Linking (e.g. entity-based search). However, for research benchmark, it is important to know which algorithms are the best fit for each of the steps in the Entity Linking workflow.

The opposite to an end-to-end evaluation is a *step-by-step strategy*. The goal of this evaluation is to provide a robust benchmark of algorithms for each step of the Entity Linking workflow. Each step is provided with the gold standard input (i.e. the correct input data for that specific step) in order to eliminate propagation of errors from previous steps. The output of each step is then evaluated separately. Despite the robustness of this approach, this type of evaluation does not account for systems that do not follow the typical Entity Linking workflow, e.g. systems that merge two steps into a single one or that create feedback loops; and it is also a highly time and labour consuming task to set up.

Finally, the *partial end-to-end* evaluation aims at evaluating the output of each Entity Linking step but by analysing the final result of the whole system. The partial end-to-end evaluation uses different metrics that are influenced only by specific parts of the Entity Linking workflow. For instance, one metric evaluates only the link between mentions and entities in the knowledge base, another metric evaluates only links with NIL, yet another one evaluates only the correct mentions recognized, whereas another metric measures the performance of the NIL Clustering.

### *2.3. Entity Linking Benchmark Initiatives*

The number of variations in Entity Linking makes it hard to benchmark Entity Linking systems. Different research communities focus on different types of text and knowledge base, and different algorithms will perform better or worse on any specific step. In this section, we present the Entity Linking benchmark initiatives to date, the Entity Linking specifications used, and the communities involved. The challenges are summarised in Table 2.3.

#### *2.3.1. TAC-KBP*

Entity Linking was first introduced in 2009 as a challenge for the Text Analysis Conference.[8] This conference was aimed at a community focused on the analysis of textual documents and the challenge itself was part of the Knowledge Base Population track (also called TAC-KBP) [45]. The goal of this track was to

[8] http://www.nist.gov/tac

| Characteristic | TAC-KBP | | | ERD | SemEval | W-NUT | NEEL | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2014 | 2015 | 2016 | 2014 | 2015 | 2015 | 2013 | 2014 | 2015 | 2016 |
| Text | newswire<br>web sites<br>discussion forum posts | | | web sites<br>search queries | technical manual<br>reports<br>formal discussions | tweets | tweets | | | |
| Knowledge Base | Wikipedia | Freebase | | Freebase | Babelnet | none | none | DBpedia | | |
| Entity | given by Type | | | given by KB | given by KB | given by Type | given by Type | | | |
| Evaluation | file | | | API | file | file | file | | API | file |
| | partial<br>end-to-end | | | end-to-end | end-to-end | end-to-end | end-to-end | | partial<br>end-to-end | |
| Target Conference | TAC | | | SIGIR | NAACL-HLT | ACL-IJCNLP | WWW | | | |

Table 1
Named Entity Recognition and Linking challenges since 2013

explore algorithms for automatic knowledge base population from textual sources. In this track, Entity Linking was perceived as a fundamental step, in which entities are extracted from text and evaluated if they already exist in the knowledge base to be populated (i.e. link to a knowledge base entry) or if they should be included in the knowledge base (i.e. link to NIL). The results of Entity Linking could be used either for direct population of knowledge bases or used in conjunction with other TAC-KBP tasks such as Slot Filling.

As of 2009, the TAC-KBP benchmark was not concerned about recognition of entities in text, in particular considering that their entities of interest were instances of types Organisation, Geo-political, and Person, and the recognition of these types of entities in text was already a well-established task in the community. The challenge was then mainly concerned with correct Entity Typing and Candidate Selection. In further years, Mention Detection and NIL Clustering were also included in the TAC-KBP pipeline [46]. Also, more entity types are now considered such as Location and Facility, as well as, multiple languages [47].

Characteristics that have been constant in TAC-KBP are the use of long textual documents, entities given by Type, and the use of encyclopedic knowledge bases. A reason for long textual documents would be that this type of text is more likely to contain contextual information to populate a knowledge base, in particular news articles and web sites. The use of entities given by Type is a direct consequence of the availability of named entity recognition algorithms based on types and the need for NIL detection. The use of an encyclopedic knowledge base was because Person, Organisation, and Geo-political entities are not domain-specific and due to the availability of Wikipedia as a free available knowledge base on the Web.

#### *2.3.2. ERD*

The Entity Recognition and Disambiguation (ERD) challenge [49] was a benchmark initiative organised in 2015 as part of the SIGIR conference[9] with the focus of enabling entity-based information retrieval. For such a task, a system needs to be able to index documents based on entities, rather than words, and to identify which entities satisfy a given query. Therefore, the ERD challenge proposed two Entity Linking tracks, a long text track based on web sites (e.g. the documents to be indexed), and a short text track based on search queries. For both tracks, entities identified in text should be linked with a subset of Freebase,[10] a large collaborative encyclopedic knowledge base containing structured data.

The Information Retrieval community, and consequently the ERD challenge, focuses on the processing of large amounts of information. Therefore, the systems evaluated should provide not only the correct results but also fulfill basic standards for large scale web systems, i.e. they should be available through Web APIs for public use, they should accept a minimum number of requests without timeout, and they should ensure a minimum uptime availability. All these standards were translated into the evaluation method of the ERD challenge that required systems to have a given Web API available for querying during the time of the evaluation. Also, large scale web systems are evaluated regarding how useful their output is for the task at hand regardless of the internal algorithms used, so the evaluation used by ERD was an end-to-end evaluation using standard information retrieval evaluation metrics (i.e. precision, recall, and f-measure).

[9] http://sigir.org/sigir2014
[10] https://developers.google.com/freebase/

#### 2.3.3. W-NUT

The community of natural language processing and computational linguistics within the ACL-IJCNLP[11] conferences have always been interested in the study of long textual documents. One of the main characteristics of these documents is that they are usually written using standard English writing. However, with the advent of Twitter and other forms of microblogging, short documents started to receive increased attention from the academic community of computational linguists in special because of their non-standard writing.

In 2015, the Workshop on Noisy-User generated text (W-NUT) [51] promoted the study of documents that are not written in standard English, with tweets as the focus of its two shared tasks. One of these tasks was targeted at the normalisation of text. In other words, expressions such as "r u coming 2" should be normalised into standard English on the form of "are you coming to". The second task proposed named entity recognition within tweets in which systems were required to detect mentions to entities corresponding to a list of ten entity types. This proposed task corresponds to the first two steps of the Entity Linking workflow: Mention Detection and Entity Typing.

#### 2.3.4. SemEval

Word Sense Disambiguation and Entity Linking are two tasks that perform disambiguation of textual documents through links with a knowledge base. Their main difference is that the former disambiguates the meaning of words with respect to a dictionary of word senses, whereas the latter disambiguates words with respect to a list of entity referents. These two tasks have been historically treated as different tasks due to the fact they require knowledge bases of a dissimilar nature. However, with the development of Babelnet, a knowledge base containing both entities and word senses, Word Sense Disambiguation and Entity Linking could be finally performed using a single knowledge base.

In 2015, a shared task for Multilingual All-Words Sense Disambiguation and Entity Linking [50] using Babelnet was proposed as part of the International Workshop on Semantic Evaluation (SemEval).[12] In this task, the goal was to create a system that could perform both Word Sense Disambiguation and Entity Linking. In word sense disambiguation, senses are anything that is available in the dictionary of senses. Therefore, in order to make the integration of the two tasks easier, it followed that entity is anything that is available in the knowledge base of entities. Also, given the complexity involved in joining the two tasks, the SemEval shared task focused on technical manuals, reports, and formal discussions which tend to follow a more rigid written structure than tweets or other forms of informal natural language text. The use of such well-written texts makes the task easier at the mention recognition level (i.e. Mention Detection), and leaves the challenge at the disambiguation level (i.e. Candidate Selection).

## 3. The NEEL Challenge Series

Named Entity Recognition and Entity Linking have been active research topics since their introduction by MUC-7 in 1997 and TAC-KBP in 2009, respectively. The main focus of these initiatives had been on long textual documents, such as news articles, or web sites. Meanwhile, microposts emerged as a new type of communication on the Social Web and have been a widespread format to express opinions, sentiments, and facts about entities. The popularisation of microposts through the use of Twitter,[13] an established platform for publication of microposts, reinforced a gap in the research of Named Entity Recognition and Entity Linking communities. The NEEL series was proposed as a benchmark initiative to fill this gap.

The evolution of the NEEL challenge followed the evolution of Entity Linking. The challenge was first held in 2013 under the name of Concept Extraction (CE) and was concerned with the detection of mentions to entities in microposts and the specification of their types. In the next year, already under the acronym of NEEL, the challenge also included linking mentions to an encyclopedic Knowledge Base or to NIL. In 2015 and 2016, NEEL was expanded to also include clustering of NIL mentions.

To propose a fair benchmark of solutions for Entity Linking with microposts, the organisation of the NEEL challenge had to make certain decisions concerning different Entity Linking components and the available strategies for evaluation, always taking into consideration the trends and needs of the research community focused on Web and microposts. In this section, we provide the motivation for these decisions. A discus-

[11] http://acl2015.org

[12] http://alt.qcri.org/semeval2015

[13] http://www.twitter.com

sion on their impact will be provided in further sections.

**Text.** The first decision that had to be taken regards the text used for the challenge. Twitter was chosen as the source of textual documents due to the fact that it is a well-known platform for microposts on the Web, and it provides a public API which makes it easy to extract microposts both for generation of the benchmark corpora and for future use of the evaluated Entity Linking systems. More information on how Twitter was used to build the NEEL corpora is presented in Section 4.

**Knowledge Base.** Despite the type of text used, it is important for an Entity Linking challenge that there is a balance between mentions linked to the Knowledge Base and mentions linked to NIL. A better balance enables a fairer evaluation, otherwise the challenge would advantage algorithms that perform one task better than the other. If the challenge is using long textual documents, the rate of the update of the Knowledge Base is less relevant because most documents likely contain a high number of mentions. These are partially new entities, that do not appear in the knowledge base; and partially old ones, that in fact are already represented in the knowledge base. However, in the case of tweets, the frequency of knowledge base updates is an important factor. Microposts are a dynamic form of communication usually dealing with recent events. If the collection of tweets is more recent than the entities in the knowledge base, the amount of NIL links is likely to be much higher than the links to entries in the Knowledge Base. Therefore, the rate in which the knowledge base is updated is an important factor for the NEEL challenge.

Taking this into account, we chose to use DBpedia [83], a structured knowledge base based on Wikipedia, mainly because it is frequently updated with entities appearing in events covered in social media. Another motivation of use DBpedia is that its format lends itself better to the task than Wikipedia itself. Each NEEL version used the latest available version of DBpedia.

**Definition of entities.** Due to the dynamic nature of microposts, the recognition of NILs was recognised as an important feature since the introduction of Entity Linking in the NEEL challenge in 2014. Due to that, but also to accommodate the participants from the Concept Extraction challenge, the definition of entities is given by type.

In 2013, the list of entity types was based on the taxonomy used in CoNLL 2003 [39]. From 2014 onwards, the NEEL Taxonomy (in Appendix) was created with the goal of providing a more fine-grained classification of entities. This would represent a vast amount of entities of interest in the context of the Web. The types of entities used and how the NEEL taxonomy was built is described in Section 4.

**Evaluation.** The evaluation is the main component of a benchmark initiative because, after all, the goal of benchmarking is to compare different systems applied to the same data by using the same evaluation metrics. There are two main decisions regarding the evaluation process. The first decision is about the format in which the results of each system are gathered (i.e. via file transfer or call to a Web API). The second decision regards how the results will be evaluated and which evaluation metrics will be applied.

The NEEL challenge has used different evaluation settings in different versions of the challenge. Each change has its own motivation, but the main focus for each of them was to provide a fair and comprehensive evaluation of the submitted systems.

The first decision regards the submission of a file or the evaluation through Web APIs. Both approaches have their advantages and disadvantages. The use of a file lowers the bar to new participants in the challenge because they do not need to develop a Web API in addition to the usual Entity Linking steps nor have to have a Web server available during the whole evaluation process. This was the proposed model for 2013, 2014, and 2016. However, during NEEL 2014, some participants suggested that the challenge should apply a blind evaluation, i.e. the participants should know the input data just at the time of the query in order to avoid common mistakes of tuning the system based on evaluation data. Therefore, in 2015 the submission of evaluation results was changed to Web API calls. The impact of this change was that few teams could not participate in the challenge, mainly because their Web server was not available during evaluation or their API did not have the correct signature. This format of evaluation also required extra effort of the organisation that had to advise participant teams that their web servers were not available. Given the amount of problems generated and no real benefit experienced, the organisation opted for going back to the transfer of files with the results of the systems as in previous years.

The second decision concerns the evaluation strategy, which impacts the metrics used and on the overall benchmark ranking. In this step, we either have the option for an end-to-end, a partial end-to-end, or a step-by-step evaluation. Borrowing from the named entity recognition community, the first two versions of the

challenge (i.e. 2013 and 2014) were based on an end-to-end evaluation. In this evaluation, standard evaluation metrics (i.e. precision, recall, and f-measure) were applied on top of the aggregated results of the system. A drawback of end-to-end evaluation is that in Entity Linking, if one step in the typical workflow does not perform well, its error will propagate until the last step. Therefore, an end-to-end evaluation will only evaluate based on the aggregated error from all steps. This was not a problem when the systems were required to perform one or two simple steps, but when the challenge starts requiring a larger number of steps then a more fine-grained evaluation is required.

A partial end-to-end strategy evaluates the output of each Entity Linking step by analysing only the final result of the system. This evaluation uses different metrics for each part of the workflow and had been successfully performed by multiple TAC-KBP versions. Therefore, due to its benefits for the research community, the partial end-to-end evaluation has also been applied in the NEEL challenge in 2015 and 2016. Furthermore, the NEEL challenge applied this strategy using the same evaluation tool as TAC-KBP [48], which aimed to enabling an easier interchange of participants between both communities.

The step-by-step evaluation has never being applied within the NEEL series. Despite its robustness by eliminating error propagation, it is very time consuming, in particular if participant systems do not implement the typical workflow. The evaluation process for each year as well as the specific metrics used will be discussed in Section 7.

**Target Conference.** The NEEL challenge keeps in mind that microposts are of interest of a broader community, composed of researchers in Natural Language Processing, Information Retrieval, Computational Linguistics, and also from a community interested on the World Wide Web. Given this, the NEEL Challenges were proposed as part of the International Workshop on Making Sense of Microposts that was held in conjunction with consecutive World Wide Web conferences.

In the next sections we will explain in detail how the NEEL challenges were organised, how the benchmark corpora were generated semi-manually, details of participant systems in each year, and the impact of each change in the participation in subsequent years.

## 4. Corpus Creation

The organisation of the NEEL challenges led to the yearly release of datasets of high value for the research community. Over the years, the datasets increased in size and coverage.

### *4.1. Collection procedure and statistics*

The initial 2013 challenge dataset contains 4,265 tweets collected from the end of 2010 to the beginning of 2011 using the Twitter firehose with no explicit hashtag search. These tweets cover a variety of topics, including comments on news and politics. The dataset was split into 66% training and 33% test.

The second 2014 challenge dataset contains 3,505 event-annotated tweets, where each entity was linked to its corresponding DBpedia URI. This dataset was collected as part of the Redites project[14] from 15th July 2011 to 15th August 2011 (31 days) comprising a set of over 18 million tweets obtained from the Twitter firehose. The 2014 dataset includes both event and non-event related tweets. The collection of event related tweets did not rely on the use of hashtags but on applying the *first story detection (FSD) algorithm* [19] and [20]. This algorithm relies on locality-sensitive hashing, which processes each tweet as it arrives in time. The hashing dynamically builds up tweet clusters representing events. Notice that hashing in this context refers to a compression methodology not to a Twitter hashtag. Within this collection, the FSD algorithm identified a series of events (stories) including the death of Amy Winehouse, the London Riots and the Oslo bombing. Since the challenge task was to automatically recognise and link named entities (to DBpedia referents), we built the challenge dataset considering both event and non-event tweets. While event tweets are more likely to contain entities, non-event tweets enabled us to evaluate the performance of the system in avoiding false positives in the entity extraction phase. This dataset was split into a training (70%) and testing (30%) sets. Given the task of identifying mentions and linking to the referent knowledge base entities in 2014, the class information was removed from the final release.

The 2015 challenge dataset extends the 2014 dataset. This dataset consists of tweets published over a longer period, between 2011 and 2013. In addition to this, we

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

| dataset | tweets | words | tokens | tokens/tweet | entities | NILs | total entities | entities/tweet | NILs/tweet |
|---|---|---|---|---|---|---|---|---|---|
| 2013 training | 2,815 | 10,439 | 51,969 | 18.46 | 2,107 | - | 3,195 | 1.88 | - |
| 2013 test | 1,450 | 6,669 | 29,154 | 20.10 | 1,140 | - | 1,557 | 1.79 | - |
| 2014 training | 2,340 | 12,758 | 41,037 | 17.54 | 1,862 | - | 3,819 | 3.26 | - |
| 2014 test | 1,165 | 6,858 | 20,224 | 17.36 | 834 | - | 1,458 | 2.50 | - |
| 2015 training | 3,498 | 13,752 | 67,393 | 19.27 | 2,058 | 451 | 4,016 | 1.99 | 0.22 |
| 2015 dev | 500 | 3,281 | 7,845 | 15.69 | 564 | 362 | 790 | 2.04 | 0.94 |
| 2015 test | 2,027 | 10,274 | 35,558 - | 17.54 | 2,122 | 1,478 | 3,860 | 2.32 | 0.89 |
| 2016 training | 6,025 | 26,247 | 100,071 | 16.61 | 3,833 | 2,291 | 8,665 | 1.43 | 0.38 |
| 2016 dev | 100 | 841 | 1,406 | 14.06 | 174 | 85 | 338 | 3.38 | 0.85 |
| 2016 test | 3,164 | 13,728 | 45,164 | 14.27 | 430* | 284* | 1,022* | $3.412^{+}$ | $0.95^{+}$ |

Table 2

General statistics of the training, dev, and test data sets. *tweets* refers to the number of tweets in the set; *words* 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 and being included in the gold standard. $^{+}$ figures refer to the 300 tweets of the gold standard.

also collected tweets from the Twitter firehose from November 2014 covering both event (such as the UCI Cyclo-cross World Cup) and non-event tweets. The dataset was split into training (58%), consisting of the entire 2014 dataset, development (8%), which enabled participants to tune their systems, and test (34%) from the newly added 2015 tweets.

The 2016 challenge dataset builds on the 2014 and 2015 datasets, and consists of tweets extracted from the Twitter firehose from 2011 to 2013 and from 2014 to 2015 via a selection of popular hashtags. This dataset was split into training (65%) consisting of the entire 2015 dataset, development (1%), and test (34%) sets from the newly collected tweets for the 2016 challenge.

Statistics describing the training, development and test sets are provided in Table 2. In all but the 2015 challenge the training datasets presented a higher rate of named entities linked to DBpedia than the development and test datasets. The percentage of tweets that mention at least one entity is 74.42% in the training, 72.96% in the test set for the 2013 dataset; 32% in the training, and 40% in the test set for the 2014 dataset; 57.83% in the training set, 77.4% in the development set, and 82.05% in the test set for the 2015 dataset; and 67.60% in the training set, 100% in the development set, and 9.35% in the test set for the 2016 dataset. The overlap of entities between the training and test data is 8.09% for the 2013 dataset, 13.27% for the 2014 dataset, 4.6% for the 2015 dataset, and 6.59% for the 2016 dataset. Following the Terms of Use of Twitter, for all the four challenge datasets, participants were only provided the tweet IDs and the annotations, the tweet text had to be mined from Twitter.

| CE 2013 | NEEL Taxonomy |
|---|---|
| MISC | Thing |
| PER | Person |
| LOC | Location |
| ORG | Organization |
| - | Character |
| - | Product |
| - | Event |

Table 3

Mapping between the taxonomy used in the first challenge of the NEEL series (left column), and the taxonomy used since the 2014 on (right column).

### *4.2. Annotation taxonomy and class distribution*

The taxonomy for annotating the entities changed from a four-class taxonomy, based on the taxonomy used in CoNLL 2003 [39], in 2013 to an extended version seven-type taxonomy, namely the NEEL Taxonomy (in Appendix), which is derived from the NERD Ontology [5]. This new taxonomy was introduced to provide a more fine-grained classification of the entities, covering names of characters, products and events. Furthermore, it is deemed to better answer the need to cope with the semantics diversity of named entities in textual documents as shown in [59]. Table 3 shows the mapping between the two classification schemes. Summary statistics of the entity types are provided in Table 4, 5, and 6 for the 2013, 2015,

and 2016 corpora respectively.[15] The most frequent entity type across all datasets is Person. This is followed by Organisation and Location in the 2013 and 2015 datasets. In the 2016 dataset the second and third most frequent types are Product and Organisation. The distributional differences between the entity types in the three sets are quite apparent, making the NEEL task challenging, particularly when tackled with supervised learning approaches.

| Type | Training | Test |
|---|---|---|
| Person | **1,722 (53.89%)** | **1,128 (72.44%)** |
| Location | 621 (19.44%) | 100 (6.42%) |
| Organisation | 618 (19.34%) | 236(15.16%) |
| Misceleneous | 233 (7.29%) | 95(6.10%) |

Table 4

Entity type statistics for the two data sets from 2013.

| 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%) |

Table 5

Entity type statistics for the three data sets from 2015.

| Type | Training | Dev | Test |
|---|---|---|---|
| Character | 63 (0.73%) | 19 (5.62%) | 57 (5.58%) |
| Event | 482 (5.56%) | 7 (2.07%) | 24 (2.35%) |
| Location | 1,868 (21.56%) | 17 (5.03%) | 43 (4.21%) |
| Organization | 1,641 (18.94%) | 33 (9.76%) | 158 (15.46%) |
| Person | **2,846 (32.84%)** | 120 (35.50%) | 337 (32.97%) |
| Product | 1,199 (13.84%) | **128 (37.87%)** | **355 (34.74%)** |
| Thing | 570 (6.58%) | 14 (4.14%) | 49 (4.79%) |

Table 6

Entity type statistics for the three data sets from 2016. The statistics of the Test set refer to the manually annotated set of tweets selected to generate the gold standard.

[15]The statistics cover the observable data in the corpora. Thus, the distributions of implicit classes in the 2014 corpus are not reported. The choice of removing the class information from the release was made on purpose because of the final objective of the task of having end-to-end solutions.

### *4.3. Annotation procedure*

In the 2013 challenge, 4 annotators created the gold standard; in the 2014 challenge a total of 14 annotators were used who had different backgrounds, including computer scientists, social scientists, social web experts, semantic web experts and natural language processing experts; in the 2015 challenge, 3 annotators generated the annotations; in the 2016 challenge, 2 experts took on the manual annotation campaign.

The annotation process for the 2013 dataset started with the unannotated corpus and it consisted of the following steps:

Phase 1. The corpus was split into four quarters, each was annotated by a different annotator.

Phase 2. For consistency checking, each annotator further checked the annotations that the other three performed to verify correctness.

Phase 3. Consensus, for the annotations without consensus, discussions between the four annotators was used to come to a final conclusion. This process resulted in resolving annotation inconsistencies.

Phase 4. Adjudication, a very small number of errors was also reported by the participants, which was taken into account in the final version of the dataset.

With the inclusion of entity links, the annotation process for the 2014, 2015 datasets was amended to consist of the following phases:

Phase 1. Unsupervised annotation of the corpus was performed, to extract potential entity mentions, candidate links to DBpedia, and in the case of 2015 challenge additionally entity types, that were used as input to the next stage. The candidates were extracted using the NERD framework [42].

Phase 2. The data set was divided into batches, with different annotators - three annotators in the 2014 challenge, and two annotators in the 2015 challenge - to each batch. In this phase annotations were performed using an annotation tool (e.g. CrowdFlower for the 2014 challenge dataset,[16] and GATE [43] for the 2015 challenge dataset[17]).

[16]For annotating the 2014 challenge dataset, we used Crowdflower with selected expert annotators rather than the crowd.

[17]For the 2015 challenge we chose GATE instead of Crowdflower, because GATE allows for the annotation of entities according to an ontology, and to compute inter-annotator agreement on the dataset.

The annotators were asked to analyse the annotations generated in Phase 1 by adding or removing entity annotations as required. The annotators were also asked to mark any ambiguous cases encountered. Along with the batches, the annotators also received the Challenge Annotation Guidelines.

Phase 3. Consistency checking, the annotators - three experts in the 2014 challenge, and a third annotator in the 2015 challenge - double-checked the annotations and generated the gold standard (for the training, development and test sets). Three main tasks were carried out here: *i)* cross-consistency check of entity types; *ii)* cross-consistency check of URIs; *iii)* resolution of ambiguous cases raised by the annotators. The annotators looped through Phase 2 and 3 of the process until the problematic cases were resolved.

Phase 4. Particular to the 2015 challenge, an unsupervised naive algorithm, based on exact matching of mention strings and their types, was used to generate an initial NIL Clustering.

Phase 5. Also in the 2015 challenge, based on the results of the naive algorithm, the third annotator manually verified all NIL clusters in order to remove links asserted to the wrong cluster, and merge clusters referring to the same entity. Special attention was paid to name variations such as acronyms, misspellings, and similar names.

Phase 6. Adjudication Phase, 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.

In the 2016 challenge, the training set was built on top of the 2014 and 2015 datasets in order to provide continuity with previous years and to build upon existing findings. The 2016 challenge used the NEEL Challenge Annotation Guidelines provided in 2015. Due to the intensity of the annotation task, 10% of the test set was annotated manually.[18] A random selection was performed while preserving the original distributions of types in the corpus by the law of large numbers [84]. The annotation process for the 2016 test set consisted of the following steps:

Phase 1. The data set was divided into 2 batches, one for each annotator. In this phase, annotations were performed using GATE. The annotators were asked to analyse the annotations generated in Phase 1 by adding or removing entity annotations as required. The annotators were also asked to mark any ambiguous cases encountered. Along with the batches, the annotators received the Challenge Annotation Guidelines.

Phase 2. Consistency checking, the two annotators checked each other annotations and generated the gold standard (for the training, development and test sets). Three main tasks were carried out here: *i)* cross-consistency check of entity types; *ii)* cross-consistency check of URIs; *iii)* resolution of ambiguous cases raised by the annotators. The annotators iterated further Phase 1 until the problematic cases were resolved.

Phase 3. Unsupervised NIL Clustering generation was performed, using a naive algorithm based on exact string matching of mention strings and their types.

Phase 4. One of the two expert annotators went through all NIL clusters in order to, where appropriate, include or exclude them from a given cluster.

Phase 5. Adjudication Phase, 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 were made as required.

The inter-annotator agreement (IAA) for the challenge datasets (2014, 2015 and 2016) is presented in Table 7.[19] We computed these values using the annotation diff tool in GATE. As the annotators are not only classifying predefined mentions but can also define different mentions, traditional IAA measures such as Cohen's Kappa are less suited to this task. Therefore, we measured the IAA in terms of precision, recall and F-measure [79].

The lessons learnt from building high quality gold standards are that the annotation process must be guided with Challenge Annotation Guidelines, at least two annotators must be involved in the annotation process to ensure consistency, and the feedback from the participants is valuable in improving the quality of

[18] The participants were asked to annotate the entire corpus of tweets.

[19] The inter-annotator agreement for the 2013 dataset could not be computed, as the challenge settings and intermediate data were lost due to lack organisation of the challenge.

| Dataset | Precision | Recall | F-measure |
|---|---|---|---|
| NEEL 2014 | 49.49% | 73.10% | 59.02% |
| NEEL 2015 | 97.00% | 98.5% | 98.00% |
| NEEL 2016 | 90.31% | 92.27% | 91.28% |

Table 7
Inter-Annotator Agreement on the challenge datasets

the datasets, providing complementary annotations to the cases found by experts. The Challenge Annotation Guidelines, written by experts, must describe the annotation task (for instance, entity types and NEEL taxonomy) through examples, and must be regularly updated during the annotation, describing special cases, issues encountered. In order to speed up the annotation process it is a good practice to employ an annotation tool. We used GATE because the annotation process was guided by a taxonomy-centric view. The annotation task took less time if the annotators shared the same background (e.g. all annotators were semantic web and natural language processing experts with experience in information extraction).

## 5. Corpus Analysis

While the main goals of the 2013-2016 challenges were the same, and the 2014-2016 corpora are largely built on top of each other, there are some differences among the datasets. In this section, we will analyse the different datasets according to the characteristics of the entities and events annotated in them. We hereby reuse measures and scripts from [78] and add a readability measure analysis of the corpora. Note that for the Entity Linking analyses, we can only compare the 2014-2016 NEEL corpora since the 2013 corpus (CE2013) does not contain entity links.

### *5.1. Entity Overlap*

Table 8 presents the entity overlap between the different datasets. Each row in the table represents the percentage of unique entities present in that dataset that are also represented in the other datasets.

### *5.2. Confusability*

We define the true confusability of a surface form $s$ as the number of meanings that this surface form can have.[20] 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 9. 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.

### *5.3. Dominance*

We define the true dominance of an entity resource $r_i$[21] for a given surface form $s_i$ 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:

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

The dominance statistics for the analysed datasets are presented in Table 10. 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 of variance for an automatic disambiguation system.

### *5.4. Summary*

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

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

[21] An entity resource is an entry in a knowledge base that describes that entity, for example `http://dbpedia.org/resource/Hillary_Clinton` is the DBpedia entry that describes the American politician Hillary Rodham Clinton.

| | NEEL 2014 | NEEL 2015 | NEEL 2016 |
|---|---|---|---|
| NEEL 2014 (2,380) | - | 1,630 (68.49%) | 1,633 (68.61%) |
| NEEL 2015 (2,800) | 1,630 (58.21%) | - | 2,800 (100%) |
| NEEL 2016 (2,992) | 1,633 (54.58%) | 2,800 (93.58%) | - |

Table 8

Entity overlap in the analysed datasets. Behind the dataset name in each row the number of unique entities present in that dataset is given. For each dataset pair the overlap is given as the number of entities and percentage (in parentheses).

| Corpus | Average | Min. | Max. | $\sigma$ |
|---|---|---|---|---|
| NEEL 2014 | 1.02 | 1 | 3 | 0.16 |
| NEEL 2015 | 1.05 | 1 | 4 | 0.25 |
| NEEL 2016 | 1.04 | 1 | 3 | 0.22 |

Table 9

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 $\sigma$ denotes the standard deviation.

| Corpus | Dominance | Max | Min | $\sigma$ |
|---|---|---|---|---|
| NEEL 2014 | 0.99 | 47 | 1 | 0.06 |
| NEEL 2015 | 0.98 | 88 | 1 | 0.09 |
| NEEL 2016 | 0.98 | 88 | 1 | 0.08 |

Table 10

Dominance stats for analysed datasets.

As the datasets are built on top of each other, they show a fair amount of overlap in entities between each other. This need not to be 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.

We experimented with various readability measures to assess the difficulty of the various tweet corpora. These measures would 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. Therefore we chose not to include those experimental results in this article.

## 6. Emerging Trends and Systems Overview

In the remainder of this analysis, we focus on two main tasks, namely Mention Detection and Candidate Selection. Thirty different approaches were applied in the four editions of the challenge since 2013. Table 11 lists all ranked teams.

### *6.1. Emerging Trends*

Whilst there are substantial differences between the proposed approaches, a number of trends can be observed in the top-performing named entity recognition and linking approaches to tweets. The main trend we observe is the large adoption of data-driven approaches: while in the first and second year of the challenge there was an extensive use of off-the-shelf approaches, the top ranking systems from 2013-2016 show a high dependence on the training data. This is not surprising, since these approaches are supervised, but this clearly defines that, to reach top performance, labeled data is necessary. Additionally, the extensive use of knowledge bases as dictionaries of typed entities and entity relation holder has dramatically affected performance over the years. This strategy overcomes the lexical limitations of a tweet and performs well on the identification of entities available in the knowledge base used as referent. A common phase in all submitted approaches is normalisation, meaning smoothing the lexical variations of the tweets and to translating them to language structures that can be better parsed by state-of-the-art approaches that expect more formal and well-formed text. Whilst the linguistic workflow favours the use of sequential solutions, Entity Recognition and Linking for tweets is proposed as joint step using large knowledge bases as referent entity directories. While knowledge bases support the linking of entities with mentions in text, they cannot support the identification of new (or emerging) entities. Ad-hoc solutions for tweets for the generation of NILs have been proposed, ranging from edit distance-based solutions to the use of Brown clustering.

| APPROACH | AUTHORS | NO.OF RUNS |
|---|---|---|
| | 2013 Entries | |
| 1 | Habib, M. et al. [6] | 1 |
| 2 | Dlugolinsky, S . et al. [7] | 3 |
| 3 | van Erp, M. et al. [8] | 3 |
| 4 | Cortis, K. [9] | 1 |
| 5 | Godin, F. et al. [10] | 1 |
| 6 | van Den Bosch, M. et al. [11] | 3 |
| 7 | Munoz-Garcia, O. et al. [12] | 1 |
| 8 | Genc, Y. et al. [13] | 1 |
| 9 | Hossein, A. [14] | 1 |
| 10 | Mendes, P. et al. [15] | 3 |
| 11 | Das, A. et al. [16] | 3 |
| 12 | Sachidanandan, S. et al. [17] | 1 |
| 13 | de Oliveira, D. et al. [18] | 1 |
| | 2014 Entries | |
| 14 | Chang, M. et al. [21] | 1 |
| 15 | Habib, M. et al. [22] | 2 |
| 16 | Scaiella, U. et al. [23] | 2 |
| 17 | Amir, M. et al. [24] | 3 |
| 18 | Bansal, R. et al. [25] | 1 |
| 19 | Dahlmeier, D. et al. [26] | 1 |
| | 2015 Entries | |
| 20 | Yamada, I. et al. [27] | 10 |
| 21 | Gârbacea, C. et al. [29] | 10 |
| 22 | Basile, P. et al. [30] | 2 |
| 23 | Guo, Z. et al. [28] | 1 |
| 24 | Barathi Ganesh, H. B. et al. [31] | 1 |
| 25 | Sinha, P. et al. [32] | 3 |
| | 2016 Entries | |
| 26 | Waitelonis, J. et. al. [37] | 1 |
| 27 | Torres-Tramon, P. et al. [36] | 1 |
| 28 | Greenfield, K. et al. [34] | 2 |
| 29 | Ghosh, S. et al. [33] | 3 |
| 30 | Caliano, D. et al. [35] | 2 |

Table 11
Per year submissions and number of runs for each team.

Between the first NEEL challenge on Concept Extraction (CE) and the 2016 edition we observe the following:

- tweet normalisation as first step of any approach. This is generally defined as preprocessing and it increases the expressiveness of the tweets, e.g. via the expansion of Twitter accounts and hashtags with the actual names of entities they represent, or with conversion of no-ASCII characters, and, generally, noise filtering;
- the 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 this particular domain;
- the use of high performing end-to-end approaches for the candidate selection. Such a methodology was further developed with the addition of fuzzy distance functions operating over ngrams and acronyms;
- the inclusion of a pruning stage to filter out candidate entities. This was presented in various approaches ranging from Learning-to-Rank to recasting the problem as a classification tasks. The latter showed best performance, holding more complexity in the definition of the feature sets;
- utilising hierarchical clustering of mentions to aggregate exact mentions of the same entity in the text and thus complementing the knowledge base entity directory in case of absence of an entity;
- a considerable decrease in off-the-shelf systems. These were popular in the first editions of NEEL, but in later editions their performance grew increasingly limited as the task became more constrained.

Table 12 provides an overview of the methods and features used in these four years, grouped according to the step involved in the workflow. In addition to the list of the steps listed in Figure 1.

### 6.2. Systems overview

Table 13 presents a description of the approaches used for Mention Detection combined with Typing. Participants approached the task using lexical similarity matchers, machine learning algorithms, and hybrid methods that combine the two. For 2013, the strategies yielding the best results where hybrid, where models relied on the application of off-the-shelf systems (e.g., AIDA [54], ANNIE [55], OpenNLP,[26] Illinois NET [56], Illinois Wikifier [57], LingPipe,[27] OpenCalais, Stanford NER [58], WikiMiner,[28] NERD [59], TwitterNLP [61], AlchemyAPI, DBpedia Spotlight, Zemanta) for both the identification of the boundaries

[26] https://opennlp.apache.org
[27] http://alias-i.com/lingpipe
[28] http://wikipedia-miner.cms.waikato.ac.nz

| Step | Method | Features | Knowledge Base | Off-the-Shelf Systems |
|---|---|---|---|---|
| Preprocessing | Cleaning, Expansion, Extraction | stop words, spelling dictionary, acronyms, hashtags, Twitter accounts, tweet timestamps, punctuation, capitalisation, token positions | - | - |
| Mention Detection | Approximate String Matching, Exact String Matching, Fuzzy String Matching, Acronym Search, Perfect String Matching, Levenshtein Matching, Context Similarity Matching, Conditional Random Fields, Random Forest, Jaccard String Matching, Prior Probability Matching | POS tags, 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[22] |
| 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 | AlchemyAPI,[23] OpenCalais,[24] Zemanta[25] |
| Candidate Selection | Distributional Semantic Model, Random Forest, RankSVM, Random Walk with Restart, Learning to Rank | gloss, contextual features, graph distance | Wikipedia, DBpedia | DBpedia Spotlight [60], AlchemyAPI, Zemanta, Babelfy [63] |
| NIL Clustering | Conditional Random Fields, Random Forest, Brown Clustering, Lack of candidate, Score Threshold, Surface Form Aggregation, Type Aggregation | POS tags, contextual words, n-grams length, predicted entity types, capitalization ratio, entity mention label, entity mention type | | |

Table 12

Map of the approaches per sub-task applied in the NEEL series of challenges from 2013 until 2016.

of the entity (mention detection) and the assignment of a semantic type (entity typing). The top performing system resulted to be System 1, which proposed an ensemble learning approach composed of a Conditional Random Fields (CRF) and a Support Vector Machines (SVM) with a radial basis function kernel specifically trained with the challenge dataset. The ensemble is performed via a union of the extraction results, while the typing is assigned via the class computed by the CRF.

The 2014 systems approached the Mention Detection task adding lexicons and features computed from DBpedia resources. System 14, the best performing system, used a matcher from ngrams computed from the text and the lexicon entries taken from DBpedia. From the 2014 on, we observe more approaches favouring recall in the Mention Detection, while focusing less on using linguistic features for mention detection. System 15, proposed by the same authors of the best performing system in 2014, addressed the Mention Detection with a large set of linguistic features and lexicon related (such as the probability of the candidate obtained from the Microsoft Web N-Gram services, or its appearance in WordNet) and using an SVM classifier with a radial basis function kernel specifically trained with the challenge data. Such an approach resulted in high precision, but it slightly penalised recall.

The 2015 best performing approach for Mention Detection, System 20, was largely inspired by the 2014 winning approach: the use of ngrams used to look

| Team | External System | Main Features | Mention Detection Strategy | Language Resource |
|---|---|---|---|---|
| | 2013 Entries | | | |
| 1 | AIDA | IsCap, AllCap, TwPOS2011 | CRF and SVM (RBF) | YAGO, Microsoft ngrams, WordNet |
| 2 | ANNIE, OpenNLP, Illinois NET, Illinois Wikifier, LingPipe, OpenCalais, StanfordNER, WikiMiner | IsCap, AllCap, LowerCase, isNP, isVP, Token length | C4.5 decision tree | Google Gazetteer |
| 3 | StanfordNER, NERD, TwitterNLP | IsCap, AllCap, Prefix, suffix, TwPOS2011, First word, last word | SVM SMO | - |
| 4 | ANNIE | IsCap, ANNIE Pos | ANNIE | DBpedia and ANNIE Gazetteer |
| 5 | Alchemy, DBpedia Spotlight, OpenCalais, Zemanta | - | Random Forest | - |
| 6 | - | PosTreebank, lowercasing | IGTree memory-based taggers | Geonames.org Gazetteer, JRC names corpus |
| 7 | Freeling | Ngram, PosFreeling 2012, isNP, Token Length | Lexical Similarity | Wiki and DBpedia Gazetteers |
| 8 | NLTK [62] | ngrams, NLTKPos | Lexical Similarity | Wikipedia |
| 9 | Babelfy API [63] | - | Lexical Similarity | DBpedia and BabelNet |
| 10 | DBpedia Spotlight | ngrams, IsCap, AllCap, lower case | CRF | DBpedia, BALIE Gazetteers |
| 11 | - | Stem, IsCap, TwPos2011, Follows | CRF | Country names, City names Gazetteers, Samsad and NICTA dictionaries, IsOOV |
| 12 | - | IsCap, prefix, suffix | CRF | Wiki and Freebase Gazetteers |
| 13 | - | ngram | PageRank, CRF | YAGO, Wikipedia, WordNet |
| | 2014 Entries | | | |
| 14 | - | ngrams, stop words removal, punctuation as tokens | Lexical Similarity | Wikipedia and Freebase lexicons |
| 15 | TwiNER [66] | Regular Expression, Entity phrases, N-gram | TwiNER and CRF | DBpedia Gazetteer, Wikipedia |
| 16 | TAGME [64] | Wikipedia anchor texts, N-grams | Collective agreement and Wikipedia statistics | Wikipedia |
| 17 | StanfordNER | - | - | NER Dictionary |
| 18 | TwitterNLP | proper nouns sequence, ngrams | - | Wikipedia |
| 19 | DBpedia Spotlight, TwitterNLP | Unigram, POS, lower, title and upper case, stripped words, isNumber, word cluster, DBpedia | CRF | DBpedia Gazetteer, Brown Clustering [65] |
| | 2015 Entries | | | |
| 20 | - | ngrams | Lexical Similarity joint with CRF, Random Forest | Wikipedia |
| 21 | Semanticizer | - | CRF | DBpedia |
| 22 | POS Tagger | ngrams | Maximun Entropy | DBpedia |
| 23 | TwitIE [67] | - | - | DBpedia |
| 24 | TwitIE | tokens | - | DBpedia |
| 25 | - | tokens | CRF joint with POS Tagger | - |
| | 2016 Entries | | | |
| 26 | - | unigrams | Lexical Similarity | DBpedia |
| 27 | GATE NLP | tokens | CRF | - |
| 28 | - | ngrams | Lexical Similarity | DBpedia |
| 29 | Stanford NER and ARK Twitter POS tagger [68] | tokens and POS | CRF | - |
| 30 | - | tokens | Lexical Similarity and Lexical Similarity | - |

Table 13

Shows per year submissions and number of runs for each team for the Mention Detection phase.

up resources in DBpedia and a set of lexical features such as POS tags and position in tweets. The type was assigned by a Random Forest classifier specifically trained with the challenge dataset and using as features linguistic features (such as POS tags, position in tweets, capitalization), DBpedia related features (such as page rank), word embeddings (contextual features), temporal popularity knowledge of an entity extracted from Wikipedia page view data, and string similarity measures to measure the similarity between the title of the entity and the mention (such as edit distance).

The 2016 best performing system, System 26, implements a lexicon matcher to match the entity in the knowledge base to the unigrams computed from the text. The approach proposed a preliminary stage of tweet normalisation resolving acronyms, hashtags to mentions written in natural language.

From 2014 on, the challenge task required participants to produce systems that were also able to link the detected mentions to their corresponding DBpedia link (if existing). Table 14 describes the approaches taken by the 2014, 2015, 2016 participants for the Candidate Detection and Selection, and NIL Clustering stages. In 2014, most of the systems proposed a Candidate Selection step as subsequent of the Mention Detection stage, thus using the output as input for finding the right link. However, the best performing system (System 14), approached the Candidate Selection as a joint stage mention detection and link assignment, proposing the so-called end-to-end system. As opposed to most of the participants which used off-the-shelf tools, System 14 proposed a SMART gradient boosting algorithm [82], specifically trained with the challenge dataset where the features are textual features (such as textual similarity, contextual similarity), graph-based features (such as semantic cohesiveness between the entity-entity and entity-mention pairs), and statistical features (such as mention popularity using the Web as archive). The majority of the systems, including System 14, applied name normalisation for feature extraction, which was useful for identifying entities originally appearing as hashtags, or username mentions. Among the most commonly used external knowledge sources are: NER dictionaries (e.g., Google CrossWiki); Knowledge Base Gazetteers (e.g., Yago, DBpedia); weighted lexicons (e.g., Freebase, Wikipedia); other sources (e.g., Microsoft Web N-gram).[29] A wide range of features were investigated for Candidate Selection strategies: ngrams, by capturing jointly the local (within a tweet) and global (within the knowledge base) contextual information of an entity via graph-based features (e.g., entity semantic cohesiveness). Other novel features included the use of Twitter account metadata and popularity-based statistical features for mentions and entity characterisation respectively.

In the 2015 challenge, System 20 (ranked first) proposed an enhanced version of the 2014 challenge winner, combined with a pruning stage meant to increase the precision of the Candidate Selection while considering the role of the type being assigned by a Conditional Random Field (CRF) classifier. In particular, System 20 is a five-sequential stage approach: preprocessing, generation of potential entity mentions, candidate selection, NIL detection, and entity mention typing. In the preprocessing stage, it is proposed a tokenisation and Part-of-Speech (POS) tagging approach based on [68], 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 one n-gram corresponds to one character). 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: 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 case of mention overlap the span with the highest score is selected. NIL clustering is addressed as a supervised learning task, in which a Random Forest classifier is used. The features consist of 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. The other approaches can be classified as sequential, where the complexity is moved to only performing the right matching of the ngram from the text and the (candidate) entity in the knowledge base. Most of these approaches exploit the popularity of the en-

[29] http://research.microsoft.com/apps/pubs/default.aspx?id=130762

tities and apply distance similarity functions to better rank entities. From the analysis, the move to controlled fully supervised in-house pipelines emerges while the use of external systems is significantly reduced. The 2015 challenge introduced the task of linking mentions to novel entities, i.e. not present in the knowledge base. All approaches in this challenge exploit lexical similarity distance functions and class information of the mentions.

In 2016, the top performing system, System 26, proposed a lexicon-based joint Mention Extraction and Candidate Selection approach, where unigrams from tweets are mapped to DBpedia entities. A preprocessing stage cleans and classifies the part-of-speech tags, and normalises the initial tweets converting alphabetic, numeric, and symbolic Unicode characters to ASCII equivalents. For every entity candidate, it considers local and context-related features. Local features include the edit distance between the candidate labels and the ngram, the candidates link graph popularity, its DBpedia type, the provenance of the label and the surface form that matches best. The context-related features assess the relation of a candidate entity to the other candidates within the given context. They include graph distance measurements, connected component analysis, or centrality and density observations using as pivot the DBpedia graph. 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.

The other approaches implement linguistic pipelines 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. While we observed a reduction in complexity for the NIL clustering, resulting in only considering the lexical distance of the mentions as for System 27 with the Monge-Elkan similarity measure [80], or System 28, that experimented the normalised Damerau-Levenshtein, performing better than Brown clustering [81].

## 7. Evaluation Strategies

In this section, the evaluation metrics used in the different challenges are described.

### *7.1. 2013 Evaluation Measures*

In 2013, the submitted systems were evaluated based on performance in extracting a mention and assigning its correct class as assigned in the Gold Standard $(GS)$. Thus a system was requested to provide a set of tuples of the form: $(m, t)$, where $m$ is the mention and $t$ is the type, which are then compared against the tuples of the gold standard $(GS)$. A type is any valid materialisation of the class defined in Table 3 and defined as Person-type, Organisation-type, Location-type, Misc-type. The precision $(P)$, recall $(R)$ and F-measure $(F_1)$ metrics were computed for each entity type. The final result for each system was reported as the average performance across the four entity types considered in the task. The evaluation was based on macro-averages across annotation types and tweets.

We performed a *strict match* between the tuples submitted and those in the GS. A *strict match* refers to an exact match, with conversion to lowercase, between a system value and the GS value for a given entity type $t$. Let $(m, t) \in S_t$ denote the set of tuples extracted for an entity type $t$ by system $S$; $(m, t) \in GS$ denotes the set of tuples for entity type $t$ in the gold standard. Then the set of true positives $(TP)$, false positives $(FP)$ and false negatives $(FN)$ for a system is defined as:

$$TP_t = \{(m,t)_S \in S | \exists (m,t)_{GS} \in GS\} \quad (1)$$

$$FP_t = \{(m,t)_S \in S | \nexists (m,t)_{GS} \in GS\} \quad (2)$$

$$FN_t = \{(m,t)_{GS} \in GS | \nexists (m,t)_S \in S\} \quad (3)$$

Since we require strict matches, a system must both detect the correct mention $(m)$ and extract the correct entity type $(t)$ from a tweet. Then for a given entity type we define:

$$P_t = \frac{|TP_t|}{TP_t \cup FP_t} \quad (4)$$

$$R_t = \frac{|TP_t|}{TP_t \cup FN_t} \quad (5)$$

| TEAM | EXTERNAL SYSTEM | MAIN FEATURES | CANDIDATE SELECTION STRATEGY | LINGUISTIC KNOWLEDGE |
|---|---|---|---|---|
| | 2014 Approaches | | | |
| 14 | - | ngrams, lower case, entity graph features (entity semantic cohesiveness), popularity-based statistical features (clicks and visiting information from the Web) | DCD-SSVM[70] and SMART gradient boosting | Wikipedia, Freebase |
| 15 | Google Search | ngrams, DBpedia and Wikipedia links, capitalisation | SVM | Wikipedia, DBpedia, WordNet, Web N-Gram, YAGO |
| 16 | TAGME | link probability, mention-link commonness distance | C4.5 (for taxonomy-filter) | Wikipedia, DBpedia |
| 17 | - | prefix, POS, suffix, Twitter account metadata, normalised mentions, trigrams | Entity Aggregate Prior, Prefix-tree Data Structure Classifier, Lexical Similarity | Wikipedia, DBpedia, YAGO |
| 18 | - | wikipedia context-based measure, anchor text measure, Twitter entity popularity | LambdaMART | Wikipedia Gazetteer, Google Cross Wiki Dictionary |
| 19 | Wikipedia Search API, DBpedia Spotlight, Google Search | mentions | Lexical Similarity and Rule-based | Wikipedia, DBpedia |
| | 2015 Approaches | | | |
| 20 | - | word embeddings, entity popularity, commonness distance, string similarity distance | Random Forest, Logistic Regression | DBpedia |
| 21 | Semanticizer | - | Learning to Rank | DBpedia |
| 22 | - | mentions | Lesk [69] | DBpedia |
| 23 | - | mentions, PageRank | Random Walks | DBpedia |
| 24 | - | mentions | Lexical Similarity | DBpedia |
| 25 | DBpedia Spotlight | mentions | Lexical Similarity | - |
| | 2016 Approaches | | | |
| 26 | - | graph distances, connected component analysis, or centrality and density observations | Learning to Rank | DBpedia |
| 27 | - | mentions, graph distances | Lexical Similarity | DBpedia |
| 28 | - | commonness, inverse document frequency anchor, term entity frequency, TCN, term entity frequency, term frequency paragraph, and redirect | SVM | DBpedia |
| 29 | Bebelfy | mentions | - | - |
| 30 | - | mentions | Lexical Similarity, context similarity | Wikipedia |

Table 14
Presents per year submissions and number of runs for each team for the Candidate Selection phase.

Then it is computed the precision and recall on a per-entity-type basis as:

$$P = \frac{P_{PER} + P_{ORG} + P_{LOC} + P_{MISC}}{4} \quad (6)$$

$$R = \frac{R_{PER} + R_{ORG} + R_{LOC} + R_{MISC}}{4} \quad (7)$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \quad (8)$$

Submissions were evaluated offline as participants were asked to annotate in a short time window a test set of the $GS$ and to send the results in a TSV[30] file.

[30]TSV stands for tab separated value.

### 7.2. 2014 Evaluation Measures

In 2014, a system $S$ was evaluated in terms of its performance in extracting both mentions and links from tweets from a set of tweets. For each tweet of this set, a system $S$ provided a tuple of the form: $(m, l)$ where $m$ is the mention and $l$ is the link. A link is any valid DBpedia URI[31] that points to an existing resource (e.g. `http://dbpedia.org/resource/Barack_Obama`). The evaluation consisted of comparing the submission entry pairs against those in $GS$. The measures used to evaluate each pair are precision ($P$), recall ($R$), and f-measure ($F_1$). The evaluation was based on micro-averages.[32]

The evaluation procedure involved an *a priori* normalisation stage for each submission. Since some DBpedia links lead to redirect pages that point to final resources, we implemented a resolve mechanism for links that was uniformly applied to all participants. In the next step, the correctness of tuples provided by a system $S$ as the exact-match of the mention and the link was assessed. Here the tuple order was also taken into account. We define $(m, l)_S \in S$ as the set of pairs extracted by the system S, $(m, l)_{GS} \in GS$ denotes the set of pairs in the gold standard. We define the set of true positives ($TP$), false positives ($FP$), and false negatives ($FN$) for a given system as:

$$TP_l = \{(m, l)_S \in S | \exists (m, l)_{GS} \in GS\} \tag{9}$$

$$FP_l = \{(m, l)_S \in S | \nexists (m, l)_{GS} \in GS\} \tag{10}$$

$$FN_l = \{(m, l)_{GS} \in GS | \nexists (m, l)_S \in S\} \tag{11}$$

$TP_l$ defines the set of relevant pairs in $S$, in other words, the set of pairs in $S$ that match the ones in $GS$. $FP_l$ is the set of irrelevant pairs in $S$, in other words the pairs in $S$ that do not match the pairs in $GS$. $FN_l$ is the set of false negatives denoting the pairs that are not recognised by $S$, yet appear in $GS$. As our evaluation is based on a micro-average analysis, we sum the individual true positives, false positives, and false negatives. As we require an exact-match for pairs $(m, l)$ we are looking for strict entity recognition and linking matches; each system has to link each recognised entity to the correct resource $l$. Precision, Recall, $F_1$ are defined as in Equation 12, Equation 13, Equation 8 respectively.

$$P = \frac{\sum_l |TP_l|}{\sum_l TP_l \cup FP_l} \tag{12}$$

$$R = \frac{\sum_l |TP_l|}{\sum_l TP_l \cup FN_l} \tag{13}$$

Submissions were evaluated offline, where participants were asked to annotate in a short time window the TS and to send the results in a TSV file.

### 7.3. 2015 and 2016 Evaluation Measures

In the 2015 and 2016 editions of the NEEL challenge, systems were evaluated according to the number of mentions correctly detected, their type correctly asserted (i.e. output of Mention Detection and Entity Typing), the links correctly assigned between a mention in a tweet and a knowledge base entry, and a NIL assigned when none knowledge base entry disambiguates the mention.

The required outputs were measured using a set of three evaluation metrics: *strong_typed_mention_match*, *strong_link_match*, and *mention_ceaf*. These metrics were combined into a final score (Equation 14).

$$\begin{aligned} score = {} & 0.4 * mention\_ceaf \\ & + 0.3 * strong\_typed\_mention\_match \\ & + 0.3 * strong\_link\_match, \end{aligned} \tag{14}$$

where the weights are empirically assigned to favour more the role of the $mention\_ceaf$, i.e. the ability of a system $S$ to link the mention either to an existing entry in DBpedia or to a NIL entry generated by $S$ and identified uniquely and consistently across different NILs.

The *strong_typed_mention_match* measures the performance of the system regarding the correct identification of mentions and their correct type assertion.

[31] We considered all DBpedia v3.9 resources valid.

[32] Since the 2014 NEEL Challenge on, we opted to weigh all instances of $TP$, $FP$, $FN$ for each tweet in the scoring, instead of weighing harmonically. This gives a better and detailed effectiveness of the system performances across different targets (typed mention, links) and tweets.

The detection of mentions is still based on strict matching as in previous versions of the challenge. Therefore true positive (Equation 9), false positive (Equation 2), and false negative (Equation 3) are still calculated in the same manner. However, the measurement of precision and recall changed slightly. In 2013, we used macro-averaged precision and recall. In this case, the impact of each mention (whether detected or not) in the final evaluation will depend on how many mentions appear in the same tweet. A wrong mention detection in a tweet with five mentions would have a smaller impact on the evaluation score than a wrong mention detected in a tweet with just one mention. In other words, for a macro-average metric, when more mentions in a tweet are present, a single mention impacts the result less. In 2015 we used micro-averaged metrics. In a micro-averaged precision and recall setup, each mention has an equal impact on the final result, regardless of how many mentions appear in the same tweet. Therefore, precision ($P$) is calculated according to the Equation 15 and recall ($R$) according to Equation 16. Finally, *strong_typed_mention_match* is the micro-averaged ($F_1$) as given by Equation 8.

$$P = \frac{\sum_t |TP_t|}{\sum_t TP_t \cup FP_t} \tag{15}$$

$$R = \frac{\sum_t |TP_t|}{\sum_t TP_t \cup FN_t} \tag{16}$$

The *strong_link_match* metric measures the correct link between a correctly recognized mention and a knowledge base entry. For a link to be considered correct, a system must detect a mention ($m$) and its type correctly ($t$) as well as the correct Knowledge Base entry ($l$). Note also that this metric does not evaluate links to NIL. The detection of mentions is still based on strict matching as in previous versions of the challenge. Therefore true positive (Equation 9), false positive (Equation 10), and false negative (Equation 11) are still calculated in the same manner. This metric is also based on micro-averaged precision and recall as defined in Equation 12 and Equation 13 and the $F_1$ as in Equation 8.

The last metric in our evaluation score is given by the Constrained Entity-Alignment F-measure (CEAF) [53]. This is a metric that measures coreference chains and is used to jointly evaluate Candidate Selection and NIL Clustering steps. Let $E = \{m_1, \dots m_n\}$ denote the set of all mentions linked to $e$, where $e$ is either a knowledge base entry or a NIL identifier. *mention_ceaf* finds the optimal alignment between the sets provided by the system and the gold standard and then performs the micro-averaged precision and recall over each mention.

In 2015, submissions were evaluated through an online process as participants were required to implement their systems as a publicly accessible web service following a REST-based protocol, where they could submit up to 10 contending entries to a registry of the NEEL challenge services. Each endpoint had a Web address (URI) and a name, which was referred as runID. Upon receiving the registration of the REST endpoint, calls to the contending entry were scheduled for two different time windows, namely, $D-Time$ - to test the APIs, and $T - Time$ - for the final evaluation and metric computations. 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 [84]. Details of the algorithm is listed in Algorithm 1. This offered the opportunity to measure the computing time systems spent in providing the answer. The computing time was proposed to solve potential draws from Equation 14.

**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**

Where $E$ is the set of entities, and $T$ is the set of tweets.

As setting up a REST API increased the system implementation load on the participants, we reverted back to an offline evaluation setup in 2016. As in previous challenges, participants were asked to annotate

the TS during a short time window and to send the results in a TSV file which was then evaluated by the challenge chairs.

### 7.4. Summary

Three editions out of four followed an offline evaluation procedure. A discontinuity was introduced in 2015 with the introduction of the online evaluation procedure. Two issues were noted by the participants of the 2015 edition: *i)* increasing complexity of the task, going beyond the pure NEEL objectives; *ii)* unfair comparison of the computing time with respect to big players that can afford better computing resources than small research teams. These motivations caused the use of a conventional offline procedure for the 2016 edition. The emerging trend sees a consolidation of a standard de-facto scorer that was proposed in TAC-KBP and also now successfully adopted and widely used in our community. This scorer allows to measure the performance of the approaches in the entire annotation pipeline, ranging from the Mention Extraction, Candidate Selection, Typing, and detection of novel entities from highly dynamic contexts such as tweets.

## 8. Results

This section presents a compilation of the NEEL challenges results across the years. As the NEEL task differs across years, the results among these years are not entirely comparable. Table 15, shows results for the NEEL 2013 task, where we report scores averaged for the four entity types analysed on this task.

The 2013 task consisted of building systems that could identify four entity types (i.e., Person, Location, Organisation and Miscellaneous) in a tweet. This task proved to be challenging, with some approaches favouring precision over recall. The best rank in precision was obtained by Team 1, which used a combination of rule types and data driven approaches achieving a 76.4% performance. For recall, results varied across the four entity types with results for the miscellaneous and organisation types ranking the lowest. Averaging over entity types, the best approach was obtained by Team 2, whose solution relied on gazetteers. All top 3 teams ranked by F-measure followed a hybrid approach combining rules and gazetteers.

The 2014 challenge task extended the concept extraction challenge by not only considering the entity type recognition but also the linking of entities to the DBpedia v3.9 knowledge base. Table 16, presents the results for this task, which follow the evaluation described in Section 7. There was a clear winner that outperformed all other systems on all three metrics and it was proposed by the Microsoft Research Lab Redmond.[33] Most of the 2014 submissions followed a sequential approach doing first the recognition and then the linking. The winning system (Team 14) introduced a novel approach, namely joint learning of recognition and linking from the training data. This approach outperformed the second best team in F-measure with over 15%.

The 2015 task extended the 2014 recognition and linking tasks with a clustering task. For this task participants had to provide clusters where each cluster

| 2013 Entries | | | |
|---|---|---|---|
| TEAM | P | R | $F_1$ |
| 1 | 0.764 | 0.604 | 0.67 |
| 2 | 0.724 | 0.613 | 0.662 |
| 3 | 0.735 | 0.611 | 0.658 |
| 4 | 0.734 | 0.595 | 0.61 |
| 5 | 0.688 | 0.546 | 0.589 |
| 6 | 0.774 | 0.548 | 0.589 |
| 7 | 0.683 | 0.483 | 0.561 |
| 8 | 0.685 | 0.5 | 0.54 |
| 9 | 0.662 | 0.482 | 0.518 |
| 10 | 0.627 | 0.383 | 0.494 |
| 11 | 0.564 | 0.43 | 0.491 |
| 12 | 0.501 | 0.468 | 0.489 |
| 13 | 0.53 | 0.402 | 0.399 |

Table 15
Scores achieved for the NEEL 2013 submissions.

| 2014 Entries | | | |
|---|---|---|---|
| TEAM | P | R | $F_1$ |
| 14 | 77.1 | 64.2 | 70.06 |
| 15 | 57.3 | 52.74 | 54.93 |
| 16 | 60.93 | 42.25 | 49.9 |
| 17 | 53.28 | 39.51 | 45.37 |
| 18 | 50.95 | 40.67 | 45.23 |
| 19 | 49.58 | 32.17 | 39.02 |

Table 16
Scores achieved for the NEEL 2014 submissions.

[33] https://www.microsoft.com/en-us/research/lab/microsoft-research-redmond/

contained only mentions to the real world entity. For 2015 we also computed the latency of each system. Table 17 presents a ranked list of results for the 2015 submissions. The last column shows the final score for each participant following Equation 14. Here the winner (Team 20) outperformed the second best with a boost in tagging $F_1$ of 41.9%, in clustering $F_1$ of 28%, and linking $F_1$ of 23.9%. Team 20 improved upon the second best team on the general score with 33.1%. For 2015, the winner team followed an End-to-End system for both candidate selection and mention typing, along with a linguistic pipeline to perform entity typing and filtering. As in 2014, the best ranked system was proposed by Studio Ousia,[34] a company focusing on knowledge extraction and artificial intelligence.

Finally, the 2016 challenge followed the same task as 2015. Team 26 outperformed all other participants, with an overall $F_1$ score of 0.5486 and a delta difference of 16.58% compared to the second-best approach. Team 26 used a learning-to-rank approach for the candidate selection task along with a series of graph-based metrics making use of DBpedia as their main linguistic knowledge source.

## 9. Conclusion

The NEEL challenge series was established in 2013 to foster the development of novel automated approaches for mining semantics from tweets and providing the community with standardised benchmark corpora, enabling the community to compare systems.

This paper describes the decisions and procedures followed in setting up and running the task. We first described the annotation procedures used to create the NEEL corpora over the years. The procedures were incrementally adjusted over time to provide continuity and ensure reusability of the approaches over the different editions. While the consolidation has provided consistent labeled data, it has also showed the robustness of the community.

We also described the different approaches proposed by the NEEL challenge participants. Over the years, we witnessed a convergence of the approaches towards data-driven solutions supported by knowledge bases. Knowledge bases are prominently used as a source to discover known entities, relations among data, and labelled data for selecting candidates and suggesting novel entities. Data-driven approaches have become, with variations, the leading solution. Despite the consolidated number of options for addressing the challenge task, the participants' results show that the NEEL task remains challenging in the microposts domain.

Furthermore, we explained the different evaluation strategies used in different challenges. These changes were driven by a desire to ensure fairness of the evaluation, transparency, and correctness. These adaptations involve the use of in-house scoring tools in 2013 and 2014, which were made publicly available and discussed in the community. Since 2015 the TAC-KBP challenge scorer was adopted to both leverage from the wide experience developed in the TAC-KBP community and to measure, while down-breaking the analysis to account for the clustering.

Thanks to the yearly releases of the annotations and tweet IDs with a public license, the NEEL corpus has started to become widely adopted. Beyond the thirty teams who completed the evaluations in four years, more than three hundred participants have contacted the NEEL organisers with a request to acquire the corpora. The teams come from more than twenty different countries and are both from academia and industry. The 2014 and 2015 winners were companies operating in the field, respectively Microsoft and Studio Ousia. The 2013 and 2016 winners were academic teams. The success of the NEEL challenges is also illustrated by the sponsorships of the challenges offered by companies (ebay[35] in 2013 and SpazioDati[36] in 2015) and research projects (LinkedTV[37] in 2014, and FREME[38] in 2016).

The NEEL challenges also triggered the interest of local communities such as the NEEL-IT. This community is pushing the NEEL guidelines (with minor variations due to the intra-language dependencies) and know-how to create a benchmark for sharing the algorithms and results of mining semantics from Italian tweets. In 2015, we also built bridges with the TAC community. We plan to strengthen these and to involve a larger audience of potential participants ranging from Linguistics, Machine Learning, Knowledge Extraction and Data and Web Science.

Future work involves the generation of corpora that account for the low variance of entity-type semantics.

[34] http://www.ousia.jp/en/

[35] http://www.ebay.com

[36] http://www.spaziodati.eu

[37] http://www.linkedtv.eu

[38] http://freme-project.eu/

| 2015 Entries | | | | | |
|---|---|---|---|---|---|
| TEAM | TAGGING $F_1$ | CLUSTERING $F_1$ | LINKING $F_1$ | LATENCY[S] | SCORE |
| 20 | 0.807 | 0.84 | 0.762 | 8.5±3.62 | 0.8067 |
| 25 | 0.388 | 0.506 | 0.523 | 0.13±0.02 | 0.4757 |
| 21 | 0.412 | 0.643 | 0.316 | 0.19±0.09 | 0.4756 |
| 22 | 0.367 | 0.459 | 0.464 | 2.03±2.35 | 0.4329 |
| 23 | 0.329 | 0.394 | 0.415 | 3.41±7.62 | 0.3808 |
| 24 | 0 | 0.001 | 0 | 12.89±27.6 | 0.004 |

Table 17

Scores achieved for the NEEL 2015 submissions. Tagging refers to *strong_typed_mention_match*, clustering refers to *mention_ceaf*, and linking to *strong_link_match*.

| 2016 Entries | | | | |
|---|---|---|---|---|
| TEAM | TAGGING $F_1$ | CLUSTERING $F_1$ | LINKING $F_1$ | SCORE |
| 26 | 0.473 | 0.641 | 0.501 | 0.5486 |
| 27 | 0.246 | 0.621 | 0.202 | 0.3828 |
| 28 | 0.319 | 0.366 | 0.396 | 0.3609 |
| 29 | 0.312 | 0.467 | 0.248 | 0.3548 |
| 30 | 0.246 | 0.203 | 0.162 | 0.3353 |

Table 18

Scores achieved for the NEEL 2016 submissions. Tagging refers to *strong_typed_mention_match*, clustering refers to *mention_ceaf*, and linking to *strong_link_match*.

We aim to create larger datasets covering a broader range of entity types and domains within the Twitter sphere. The 2015 enhancements in the evaluation strategy, which accounts for computational time, highlighted new challenges when focusing on an algorithm's efficiency vs efficacy. Since more efforts on handling large scale data mining involve distributed computing and optimisation, we aim to develop new evaluation strategies. These strategies will ensure the fairness of the results when asking participants to produce large scale annotations in a small window of time. Finally, given the increasing interest in adopting the NEEL guidelines in creating corpora for other languages, we aim to develop a multilingual NEEL challenge as a future activity.

## Acknowledgments

This work was supported by the FREME project (GA no. 644771) and by the CLARIAH-CORE project financed by the Netherlands Organisation for Scientific Research (NWO).

## References

[1] A. E. Cano Basave, A. Varga, M. Rowe, M. Stankovic, A. Dadzie, *Making Sense of Microposts (#MSM2013) Concept Extraction Challenge*, Making Sense of $3^{rd}$ Workshop on Making Sense of Microposts (#Microposts2013), 2013.

[2] A. E. Cano Basave, G. Rizzo, A. Varga, M. Rowe, M. Stankovic, A. Dadzie, *Making Sense of Microposts (#Microposts2014) Named Entity Extraction & Linking Challenge*, $4^{th}$ Workshop on Making Sense of Microposts, 2014.

[3] G. Rizzo, A. E. Cano Amparo, B. Pereira, A. Varga, *Making sense of Microposts (#Microposts2015) Named Entity rEcognition & Linking Challenge*, $5^{th}$ International Workshop on Making Sense of Microposts, 2015.

[4] G. Rizzo, M. van Erp, J. Plu, R. Troncy, *Making Sense of Microposts (#Microposts2016) Named Entity rEcognition and Linking (NEEL) Challenge*, $6^{th}$ International Workshop on Making Sense of Microposts, 2016.

[5] G. Rizzo, R. Troncy, S. Hellmann, M. Brummer, *NERD meets NIF: Lifting NLP Extraction Results to the Linked Data Cloud*, (WWW'12) Linked Data on the Web (LDOW'12), 2012

[6] M. Habib, M. Van Keulen, Z. Zhu, *Concept Extraction Challenge: University of Twente at #MSM2013*, Concept Extraction Challenge at the Workshop on Making Sense of Microposts, 2013.

[7] S. Dlugolinský, Peter Krammer, Marek Ciglan, Michal Laclavik, *MSM2013 IE Challenge: Annotowatch*, Concept Extraction Challenge at the Workshop on Making Sense of Microposts, 2013.

[8] M. Van Erp, G. Rizzo, R. Troncy, *Learning with the Web: Spotting Named Entities on the Intersection of NERD and Machine Learning*, Concept Extraction Challenge at the Workshop on Making Sense of Microposts, 2013.

[9] K. Cortis, *ACE: A Concept Extraction Approach using Linked Open Data*, Concept Extraction Challenge at the Workshop on Making Sense of Microposts, 2013.

[10] F. Godin, P. Debevere, E. Mannens, W. De Neve, R. Van de Walle, *Leveraging Existing Tools for Named Entity Recognition in Microposts*, Concept Extraction Challenge at the Workshop on Making Sense of Microposts, 2013.

[11] A. van Den Bosch, T. Bogers, *Memory-based Named Entity Recognition in Tweets*, Concept Extraction Challenge at the Workshop on Making Sense of Microposts, 2013.

[12] Ó. Muñoz-García, A. García-Silva, Ó. Corcho, *Towards Concept Identification using a Knowledge-Intensive Approach*, Concept Extraction Challenge at the Workshop on Making Sense of Microposts, 2013.

[13] Y. Genc, W. Mason, J. V. Nickerson, *Classifying Short Messages using Collaborative Knowledge Bases: Reading Wikipedia to Understand Twitter*, Concept Extraction Challenge at the Workshop on Making Sense of Microposts, 2013.

[14] A. Hossein Jadidinejad, *Unsupervised Information Extraction using BabelNet and DBpedia*, Concept Extraction Challenge at the Workshop on Making Sense of Microposts, 2013.

[15] P. Mendes, D. Weissenborn, C. Hokamp, *DBpedia Spotlight at the MSM2013 Challenge*, Concept Extraction Challenge at the Workshop on Making Sense of Microposts, 2013.

[16] A. Das, U. Burman, B. Ar, S. Bandyopadhyay, *NER from Tweets: SRI-JU System MSM 2013*, Concept Extraction Challenge at the Workshop on Making Sense of Microposts, 2013.

[17] S. Sachidanandan, P. Sambaturu, K. Karlapalem, *NERTUW: Named Entity Recognition on Tweets using Wikipedia*, Concept Extraction Challenge at the Workshop on Making Sense of Microposts, 2013.

[18] D. de Oliveira, A. Laender, A. Veloso, A. Da Silva, *FilterStream Named Entity Recognition: A Case Study at the #MSM2013 Concept Extraction Challenge*, Concept Extraction Challenge at the Workshop on Making Sense of Microposts, 2013.

[19] M. Osborne, S. Petrovic, R. Mccreadie, C. Macdonald and I. Ounis, *Bieber no more: First story detection using twitter and wikipedia*, (SIGIR'2012) Workshop on Time-aware Information Access, 2012.

[20] S. Petrovic, M. Osborne, V. Lavrenko, *Streaming First story detection with application to Twitter*, North American Chapter of the Association for Computational Linguistics (NAACL), 2010.

[21] M. Chang, B. Hsu, H. Ma, R. Loynd, K. Wang, *E2E: An End-to-End Entity Linking System for Short and Noisy Text*, $4^{th}$ International Workshop on Making Sense of Microposts (#Microposts), 2014.

[22] M. B. Habib, M. van Keule, Z. Zhu, *Named Entity Extraction and Linking Challenge: University of Twente at #Microposts2014*, $4^{th}$ International Workshop on Making Sense of Microposts (#Microposts), 2014.

[23] U. Scaiella, M. Barbera, S. Parmesan, G. Prestia, E. Del Tessandoro, M. Veri, *DataTXT at #Microposts2014 Challenge*, $4^{th}$ International Workshop on Making Sense of Microposts (#Microposts), 2014.

[24] M. Amir Yosef, J. Hoffart, Y. Ibrahim, A. Boldyrev, G. Weikum, *Adapting AIDA for Tweets*, $4^{th}$ International Workshop on Making Sense of Microposts (#Microposts), 2014.

[25] R. Bansal, S. Panem, P. Radhakrishnan, M. Gupta, V. Varma, *Linking Entities in #Microposts*, $4^{th}$ International Workshop on Making Sense of Microposts (#Microposts), 2014.

[26] D. Dahlmeier, N. Nandan, W. Ting, *Part-of-Speech is (almost) enough: SAP Research & Innovation at the #Microposts2014 NEEL Challenge*, $4^{th}$ International Workshop on Making Sense of Microposts (#Microposts), 2014.

[27] I. Yamada, H. Takeda, Y. Takefuji, *An End-to-End Entity Linking Approach for Tweets*, $5^{th}$ International Workshop on Making Sense of Microposts (#Microposts), 2015.

[28] Z. Guo, D. Barbosa, *Entity Recognition and Linking on Tweets with Random Walks*, $5^{th}$ International Workshop on Making Sense of Microposts (#Microposts), 2015.

[29] C. Gârbacea, D. Odijk, D. Graus, I. Sijaranamual, M. de Rijke, *Combining Multiple Signals for Semanticizing Tweets: University of Amsterdam at #Microposts2015*, $5^{th}$ International Workshop on Making Sense of Microposts (#Microposts), 2015.

[30] P. Basile, A. Caputo, G. Semeraro, F. Narducci, *UNIBA: Exploiting a Distributional Semantic Model for Disambiguating and Linking Entities in Tweets*, $5^{th}$ International Workshop on Making Sense of Microposts (#Microposts), 2015.

[31] H. B. Barathi Ganesh, N. Abinaya, M. Anand Kumar, R. Vinaykumar, K.P. Soman, *AMRITA - CENNEEL: Identification and Linking of Twitter Entities*, $5^{th}$ International Workshop on Making Sense of Microposts (#Microposts), 2015.

[32] P. Sinha, B. Barik, *Named Entity Extraction and Linking in #Microposts*, $5^{th}$ International Workshop on Making Sense of Microposts (#Microposts), 2015.

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

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

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

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

[37] J. Waitelonis, H. Sack, *Named Entity Linking in #Tweets with KEA*, $6^{th}$ International Workshop on Making Sense of Microposts (#Microposts), 2016.

[38] R. Grishman, B. Sundheim, *Message Understanding Conference-6: a brief history*, $16^{th}$ International Conference on Computational linguistics (COLING), 1996.

[39] E. F. Tjong Kim Sang, F. D. Meulder, *Introduction to the CoNLL-2003 Shared Task: Language Independent Named Entity Recognition*. $17^{th}$ Conference on Computational Natural Language Learning (CoNLL), 2003.

[40] R. C. Bunescu, M. Pasca, *Using Encyclopedic Knowledge for Named entity Disambiguation*, European Chapter of the Association for Computational Linguistics (EACL), 2006.

[41] S. Cucerzan, *Large-Scale Named Entity Disambiguation Based on Wikipedia Data*, Empirical Methods in Natural Language

Processing (EMNLP-CoNLL), 2007.

[42] G. Rizzo, M. van Erp, R. Troncy, *Benchmarking the extraction and disambiguation of named entities on the semantic web*, $9^{th}$ International Conference on Language Resources and Evaluation, 2014.

[43] 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, *Text Processing with GATE (Version 6)*, GATE, 2011.

[44] N. chinchor, P. Robinson. *MUC-7 named entity task definition*, $7^{th}$ Conference on Message Understanding, 1997.

[45] P. McNamee, H. T. Dang, *Overview of the tac 2009 knowledge base population track*, Text Analysis Conference (TAC), 2009.

[46] H. Ji, R. Grishman, H. T. Dang, *Overview of the TAC2011 Knowledge Base Population Track*, Text Analysis Conference (TAC), 2011.

[47] H. Ji, J. Nothman, B. Hachey, R. Florian, *Overview of TAC-KBP2015 Tri-lingual Entity Discovery and Linking*, Text Analysis Conference (TAC), 2015.

[48] B. Hachey, J. Nothman, W. Radford (2014), , Association for Computational Linguistics Conference (ACL), 2014.

[49] D. Carmel, M.-W. Chang, E. Gabrilovich, B.-J. P. Hsu, K. Wang, *ERD'14: Entity Recognition and Disambiguation Challenge*, Research and Development in Information Retrieval Conference (ACM SIGIR), 2014.

[50] A. Moro, R. Navigli, *SemEval-2015 task 13: multilingual all-words sense disambiguation and entity linking*, International Workshop on Semantic Evaluation (SemEval), 2015.

[51] T. Baldwin, Y.-B. Kim, M. C. de Marneffe, A. Ritter, B. Han, W. Xu, *Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition*, $53^{rd}$ Annual Meeting of the Association for Computational Linguistics and The $7^{th}$ International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP), 2015.

[52] A. Bagga, B. Baldwin, *Algorithms for scoring coreference chains*, $1^{st}$ international conference on language resources and evaluation workshop on linguistics coreference, 1998.

[53] X. Luo, *On coreference resolution performance metrics*, Human Language Technology and Empirical Methods in Natural Language Processing Conference (HLT-EMNLP), 2005.

[54] J. Hoffart, M. Amir Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, G. Weikum, *Robust Disambiguation of Named Entities in Text*, Empirical Methods in Natural Language Processing Conference (EMNLP), 2011.

[55] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*, $40^{th}$ Anniversary Meeting of the Association for Computational Linguistics (ACL), 2002.

[56] L. Ratinov, D. Roth, *Design challenges and misconceptions in named entity recognition*, $13^{th}$ Conference on Computational Natural Language Learning (CoNLL), 2009.

[57] L. Ratinov, D. Roth, D. Downey, and M. Anderson, *Local and global algorithms for disambiguation to wikipedia*, $49^{th}$ Annual Meeting of the Association for Computational Linguistics (ACL), 2011.

[58] J. R. Finkel, T. Grenager, and C. Manning, *Incorporating non-local information into information extraction systems by gibbs sampling*, $43^{rd}$ Annual Meeting on Association for Computational Linguistics (ACL), 2005.

[59] G. Rizzo, R. Troncy, *NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Web Extraction Tools*, European chapter of the Association for Computational Linguistics (EACL), 2012.

[60] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, *DBpedia spotlight: shedding light on the web of documents*, $7^{th}$ International Conference on Semantic Systems (I-Semantics), 2011.

[61] A. Ritter, S. Clark, Mausam, and O. Etzioni, *Named entity recognition in tweets: An experimental study*, Conference on Empirical Methods on Natural Language Processing (EMNLP), 2011.

[62] E. Loper, S. Bird, *NLTK: The Natural Language Toolkit*, (ACL) Workshop on Elective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, 2002.

[63] A. Moro, A. Raganato, R. Navigli, *Entity Linking meets Word Sense Disambiguation: a Unified Approach*, Transactions of the Association for Computational Linguistics (TACL), 2014.

[64] P. Ferragina, U. Scaiella, *TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities)*, $19^{th}$ International Conference on Information and Knowledge (CIKM), 2010.

[65] P. Liang, *Semi-Supervised Learning for Natural Language*, MIT, 2005.

[66] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B. S. Lee, *TwiNER: Named entity recognition in targeted Twitter stream*, $35^{th}$ International Conference on Research and Development in Information Retrieval (ACM SIGIR), 2012.

[67] K. Bontcheva, L. Derczynski, A. Funk, M.A. Greenwood, D. Maynard, N. Aswani, *TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text*, International Conference on Recent Advances in Natural Language Processing, RANLP, 2013.

[68] K. Gimpel, N. Schneider, B. O'Connor, D. Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and N. A. Smith, *Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments*, $49^{th}$ Annual Meeting of the Association for Computational Linguistics (ACL), 2011.

[69] P. Basile, A. Caputo, and G. Semeraro, *An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model*, $25^{th}$ International Conference on Computational Linguistics (COLING), 2014.

[70] M. Chang, W. Yih, *Dual Coordinate Descent Algorithms for Efficient Large Margin Structured Prediction*, Transactions of the Association for Computational Linguistics, 2013.

[71] Carl-Hugo Björnsson, *Läsbarhet*, Liber, 1968.

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

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

[74] Robert Gunning, *The Technique of Clear Writing*, McGraw-Hill, 36–37 , 1952.

[75] 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.

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

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

[78] Marieke van Erp, Pablo Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, Giuseppe Rizzo, and Joerg Waitelonis, *Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job*, $10^{th}$ International Conference on Language Resources and Evaluation (LREC), 2016.

[79] H. Cunningham, D. Maynard, K. Bontcheva, and et al., *Developing Language Processing Components with GATE Version 8 (a User Guide)*, The University of Sheffield, Department of Computer Science, 2014.

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

[81] P. F. Brown, P. V. de Souza, R. L. Mercer, V. J. Della Pietra, J. C. Lai, *Class-based N-gram Models of Natural Language*, Computation Linguistics, 18:4, 467–479, 1992.

[82] J. H. Friedman. *Greedy function approximation: A gradient boosting machine*, Annals of Statistics, 1999.

[83] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P.N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, C. Bizer, *DBpedia âĂŞ a large-scale, multilingual knowledge base extracted from Wikipedia*, Semantic Web, 6(2), 167–195, 2015

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

[85] F. Abel, I. Celik, G. Houben, P. Siehndel *Leveraging the semantics of tweets for adaptive faceted search on twitter*, International Semantic Web Conference (ISMC), 2011.

[86] A. Varga, A.E. Cano, M. Rowe, F. Ciravegna, Y. He *Leveraging the semantics of tweets for adaptive faceted search on twitter*, Web Semantics: Science, Services and Agents on the World Wide Web, 2014.

[87] F. Abel, Q. Gao, G. Houben, K. Tao *Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web*, Extended Semantic Web Conference (ESWC), 2011.

[88] P. Basile, A. Caputo, A. L. Gentile, G. Rizzo, *Overview of the EVALITA 2016 Named Entity rEcognition and Linking in Italian Tweets (NEEL-IT) Task*, $5^{th}$ Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA), 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]
```