Undefined 1 (2009) 1–5
IOS Press

Benchmark Corpora and Emerging Trends in Mining Semantics from Tweets

The NEEL Challenge: The Story So Far

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. Email: bianca.pereira@insight-centre.org*
[c] *The Content Group, Godalming, UK. Email: varga.andy@gmail.com*
[d] *Vrije Universiteit Amsterdam, Netherlands. Email: marieke.van.erp@vu.nl*
[e] *Aston Business School, Aston University. Email: ampaeli@gmail.com*

Abstract. The large number of tweets generated daily has provided means for policy makers to get 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 a series of algorithms aimed at the automatic extraction of semantics in tweets and their link to machine readable resources. While a tweet is shallowly comparable to any other textual content, it hides a complex and challenging structure featured by acronyms, abbreviations, emojis, typos, and a rich set of metadata based on entities. The NEEL series of challenges, established in 2013, has contributed to collect the emerging trends in the field and define standardized benchmark corpora for entity recognition and linking in tweets, ensuring high quality labeled data that enables easier comparisons between different approaches. This paper reports on the findings and lessons learned through an analysis of specific characteristics of the created corpora and highlighting limitations, lessons learned from the different participants in the challenges and providing guidance to implement top performing approaches in 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 proven to be useful in different applications and contexts such as music recommendation, spam detection, emergency response, market analysis, and decision making. While they are potentially gold mine, the volume of tweets generated has thus moved the focus to automatic processing. A commonly used approach for automatically extracting information from tweets is the use of textual cues, which provide contextual features for the underlying tweet content.

With the ever increasing importance of tweets in our daily life to communicate feelings, tell stories, report facts, or advertise events and products, industry and research communities have started to investigate automatic approaches to identify *named entities*. A named entity is used in the general sense of being, not requiring a material existence but requiring to be an instance of a taxonomy class. Thus, a mention to an entity in a tweet is seen as proper noun or an acronym referring to a real entity. The extent of an entity is the entire string representing the name, excluding the preceding defi-

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

nite article (i.e. "the") and any other pre-posed (e.g. "Dr","Mr") or post-posed modifiers.[1]

However, the identification, classification and linking of named entities have proven to be challenging tasks due to, among other things, the inherent characteristics of this type of data: *i)* the restricted length and *ii)* the noisy lexical nature, where 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 has established a yearly challenge in the field covering newswire, websites, discussion forum posts, ERD [41] with search queries content, and SemEval with technical manuals and reports.

The NEEL series of challenges, 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 '14, prompted the Named Entity rEcognition and Linking (NEEL) Challenge in [3]. In 2015, furthering the role of the named entity type in the process was investigated, as well as the identification of named entities that cannot be grounded because they do not have a knowledge base referent. The English DBpedia 2014 dataset was the designated reference 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, while opting for not weighting anymore the computing time and thus measuring offline the systems' performance.

The four challenges have published four different labeled corpora. The creation of the corpora followed rigid designations and protocols, this is to grant high quality labeled data that can be used as seeds for any reasoning and supervised approaches. Despite the protocols, the corpora have strengths and weaknesses that we have learned in these years.

The purpose of each challenge was also to set up an open and competitive environment that would encourage participants to deliver novel approaches or improve on existing approaches for recognizing and linking entities from tweets to either a reference knowledge base entry or NIL where such a reference does not exist. From the first (in 2013) through to the 2016 NEEL challenge, thirty research teams have submitted at least an entry to the competitions proposing state-of-the-art approaches.

More than three hundred teams have explicitly[2] acquired the corpora in the four years, underlining the importance of the challenges in the field.

The NEEL challenges have also experienced a strong involvement of the industry as both participants and funding agencies. In particular, the last aspect has generated a yearly prize assigned to the winner of the challenge. These sponsorships are testament to the growing interest in challenges related to automatic approaches for gleaning information from (the very large amounts of) social media data generated across all aspects of life, and whose knowledge content is recognised to be of value to industry.

This paper reports on the findings and the lessons learned of the last four years of the NEEL Challenge, analyzing in details the corpora, highlighting their limitations, while providing guidance to implement top performing approaches in the field from the lessons learned of 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 the field of entity identification, recognition and linking and underline the reason that has prompted the need to establish the NEEL series of challenges. We then detail the steps followed in generating the four different corpora in Section 3, followed by a thorough quantitative and qualitative analysis of the corpora in Section 4. We then list the different approaches presented and narrow down the emerging trends in Section 5, grounding the trends according to

[1] As defined in the "NEEL Challenge, Annotation Guidelines".

[2] This figure does not consider the teams who experimented with the corpora out of the challenges' timeline.

the evaluation strategies being presented in Section 6. Section 7 concludes with the lessons learned and discusses future endeavors.

2. Task 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) [39]. 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 Organization. Since then, named entities have become a key aspect in different research domains.

Having the entity recognized in a textual document was the first big challenge, but having crossed this barrier, the research community moved into a second and more challenging problem: the disambiguation of 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 more than another million people called Paul around the world. In the same manner, *Copacabana* can be a mention to the beach in Rio de Janeiro, Brazil, or to the beach in Dubrovnik, Croatia. The problem of ambiguity is translated into the question of "what is the exact entity that is mentioned in text?". To solve this problem, recognizing the mention to an entity in text is just the first step, the next one is to link the mention to an unambiguous representation of the same entity in a knowledge base. This task became known in the community as Entity Disambiguation.

The Entity Disambiguation task popularized after Bunescu and Pasca [35] in 2006 explored the use of an encyclopedia as source for entities. In particular, after Cucerzan [36] showed the benefit of using Wikipedia,[3] a free crowd-sourced encyclopedia, for such purpose. The reason why encyclopedic knowledge is so important is that an encyclopedia source contains a unique representation for each entity, along with its description, and it covers entities in a variety of domains of knowledge. 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 for entities.

In 2009, the TAC-KBP challenge [40] demonstrated a new problem to both the Entity Recognition and Entity Disambiguation communities. In Entity Recognition, the mention is recognised in text without knowing the exact entity that is being referred by the mention. In Entity Disambiguation, there was a focus only in the entities that are present in a provided knowledge base. TAC-KBP pointed out the problem that a mention identified in text by a Named Recognition strategy for instance, 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, to indicate 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 because Entity Disambiguation can be seen as a special case of Entity Linking.

Despite all efforts in the research community led by MUC-7 and TAC-KBP challenges, their unique focus was in long textual documents, such as news articles, websites, and forum discussions. Meanwhile, a new type of communication emerged on the Social Web. The creation of micro-posts in general and its popularization through the use of Twitter, as established platform for publication of micro-posts, created a gap in the research of Entity Recognition and Entity Linking communities. Therefore, in 2013, the Concept Extraction challenge was created to fill this gap through the performance of Entity Recognition in tweets. One year later, the NEEL challenge was proposed to perform not only Entity Recognition but also Entity Linking.

The evolution of the NEEL challenge, as we will describe in the remainder of this paper, followed the evolution of the Entity Linking topic. The first version of the challenge was concerned only with the recognition of entities in tweets. The year after, already under the name of NEEL, the challenge also included linking mentions to an encyclopedic knowledge or to NILs. And in 2015 and 2016, NEEL was expanded, following recent trends in the research community, to include the clustering of NIL mentions.

Since then, other challenges have been proposed for different research communities. TAC-KBP has been part of multiple Text Analysis conferences,[4] therefore its main target has been the Natural Language Processing community. In 2014, the Entity Resolution and

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

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

Disambiguation (ERD) challenge [41] was proposed as part of the SIGIR conference,[5] aiming to introduce the Entity Linking task for the Information Retrieval community. In the next year, The SemEval[6] challenge launched the Multilingual All-Words Sense Disambiguation and Entity Linking task [42] extending Entity Linking to the Computational Linguistics community, presented at the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT).[7] Still in 2015, the W-NUT challenge [43] proposed a shared task for Twitter Lexical Normalization and Named Entity Recognition aiming to introduce the task to the Computational Linguistics and Natural Language Processing communities at the ACL-IJCNLP conference.[8] The NEEL challenge was organized keeping in mind that micro-posts are of interest of a broader community composed of researches in Natural Language Processing, Information Retrieval, Computational Linguistics, and also from a community interested on the World Wide Web. Given this, the World Wide Web Conference[9] has been the target for the NEEL challenge.

From the Entity Linking challenges proposed since 2009, NEEL has a very specific niche: the micro-posts community. Each version of the NEEL challenge follows the needs and trends in the micro-posts community, and incorporates the successful characteristics of other Entity Linking challenges. In order to provide a better perspective on where the NEEL challenge positions in the range of Named Entity Recognition and Linking challenges, we will characterise well-known challenges developed since the first creation of the Concept Extraction challenge in 2013, regarding the preparation of the challenges and the evaluation process, and how NEEL differs in each of the features analysed.

2.1. *Overview of Named Entity Recognition and Entity Linking challenges*

Each Named Entity Recognition and Linking challenge was set up differently. The type of entities recognised may differ as well as the type of text, the knowledge base used, the evaluation metrics, the process used to develop the annotated corpus for evaluation, and how the results for evaluation are received by the organisation of the challenge. A summary of challenges since the Concept Extraction challenge in 2013 and their characteristics is presented in Table 2.1.

Regarding the type of textual document, NEEL and W-NUT Shared Task focus on micro-posts, in particular tweets. However, NEEL focuses on both Named Entity Recognition and Linking, whereas W-NUT focuses on Twitter Normalization and Named Entity Recognition. TAC-KBP, ERD and SemEval focus in performing Entity Linking in long textual documents, such as web sites and discussion forums. ERD also contains a short text track, however the target is search queries which have a different nature than micro-posts.

Despite of the type of text, it is important for an Entity Linking challenge to balance between mentions linked to a knowledge base and mentions linked to NIL. The better this balance the fairest is the evaluation, otherwise the challenge would give advantage to algorithms that perform only one of these tasks better. If the challenge is using long textual documents, the rate of the update of the knowledge base is not very relevant because it is very likely that each document will contain a large number of mentions, partially new entities, that do not appear in the referent knowledge base; and partially old ones, that in fact are already represented in the knowledge base. In the case of tweets, however, how often the knowledge base is updated is an important factor. Tweets are short, thus contain few mentions, and deal with recent events. If the collection of tweets is more recent than the entities in the referent knowledge base, the amount of NILs 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.

TAC-KBP and ERD Challenges focused on (static) snapshots of encyclopedic knowledge bases for comparability of the results. TAC-KBP focused on a 2008 version of Wikipedia until TAC-KBP 2014 and changed to a Freebase snapshot[10] in 2015, whereas ERD also used Freebase.[11] SemEval decided to use Babelnet[12] given its focus in word meanings as well as entities. In NEEL, we chose to use DBpedia,[13] a

[5] http://sigir.org/sigir2014/
[6] http://alt.qcri.org/semeval2015/
[7] http://naacl.org/naacl-hlt-2015/
[8] http://acl2015.org/
[9] http://www.iw3c2.org/

[10] http://basekb.com/
[11] Note that Freebase is also officially not-longer maintained knowledge base.
[12] http://babelnet.org/
[13] http://wiki.dbpedia.org/

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

Table 1
Named Entity Recognition and Linking challenges after 2013.

knowledge base based on Wikipedia data, mainly because it is frequently updated with entities appearing in events covered in social media, but also because it is in an easier format to process than Wikipedia itself. Each NEEL version used the latest available version of DBpedia.

Another aspect of Entity Linking is the definition of entity. Each challenge works with its own definition. SemEval considers anything in the knowledge base as synsets (i.e. word meanings) and possible targets for linking with mentions. In other words, entity is anything which is represented in the knowledge base. ERD also defines entities given by the knowledge base used, however, the knowledge base has been filtered regarding the types of entities.[14] TAC-KBP opted for a different approach. Its definition of entities is based on the type of each entity regardless if they appear in the knowledge base. TAC-KBP focused originally in linking only Person, Organization, and Geo-Political entities. In 2015 this focus has been expanded to cover also Facilities and Natural Locations. W-NUT also focused on a list of 10 types of entities to be recognized in tweets. In NEEL, the organisation also opted for describing the types of entities to be linked, mainly to avoid entities that are not of interest to applications consuming tweets (e.g. letters from the alphabet), but also partially to accommodate the participants coming from the Concept Extraction challenge. The types of entities were given by an ontology of entities which will be described in details in Section 3.

The format of the evaluation also varies. The results reported by the participants can be either sent in a document in a specific format or queried through a Web API. TAC-KBP, SemEval, and W-NUT release the textual documents to be annotated and require a file providing all the inferred links. On the contrary, ERD evaluates all systems by querying their APIs with the texts to be annotated as input. In NEEL, we experimented with both approaches. The 2013, 2014, 2016 versions were focused on the use of a file, while in 2015 we opted for the API. Both approaches have their advantages and disadvantages. The use of a file makes the entrance of new participants in the challenge easier because they do not need to develop a Web API in addition to the usual Entity Linking steps nor have to have an available server to be queried. However, during NEEL 2014, the participants suggested that the evaluation should be 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.

[14] The types filtered out were not disclosed in the ERD report.

2.2. The evaluation process

The goal of research challenges is to contribute to a research community through the proposal of interesting problems and the benchmarking of solutions to these problems. The proposal of new problems fosters innovation within the community and the benchmarking facilitates the discovery of state-of-the-art solutions by new entrants. Given the second contribution, the evaluation step has a central role in any challenge of measuring how performing is the winner approach.

Evaluation metrics play an important role in the evaluation process because they define the ranking of participants and inform which approach performs best for the proposed problem. In order to discuss the evaluation process and evaluation metrics, first we need to delineate the steps of the Entity Linking workflow and define the expected output of each step.

In Figure 1 we present a typical Entity Linking workflow composed of the following steps: Mention Detection, Entity Typing, Candidate Detection, Candidate Selection, and NIL Clustering. Note that, al-

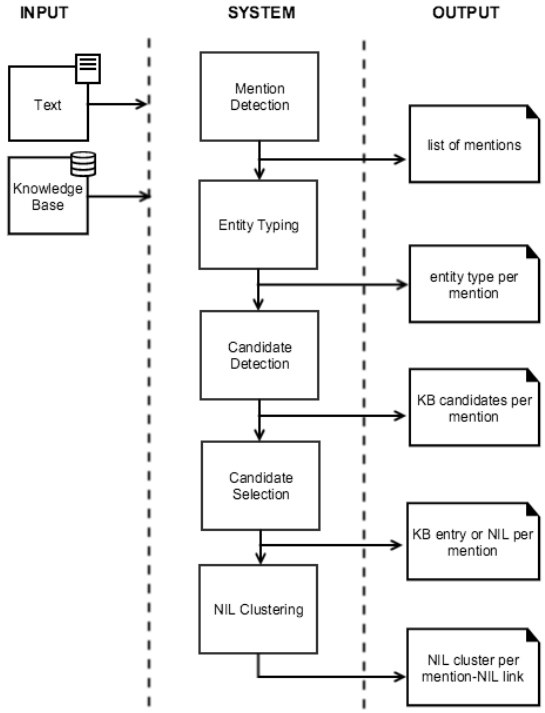

Fig. 1. Typical Entity Linking workflow with expected output of each step.

though 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 outputs all terms in text used as mentions to entities. The Entity Typing step receives mentions identified in text and outputs their type. Candidate Detection receives the same mentions as input and identifies all entries in the knowledge base that are candidates to be linked with each mention. Further, the Candidate Selection step receives mentions and their candidates list as input and outputs the correct link for each mention, either an entity in the knowledge base or a NIL. Last, the NIL Clustering receives a series of mentions linked to NIL as input and outputs clusters on mentions for 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.

When an evaluation is performed end-to-end it means that only the final result of the whole Entity Linking process is evaluated. If one step in the workflow does not perform well, then the 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 type of evaluation is used to evaluate the system as a whole without concerning with the algorithms used within the system. The end-to-end evaluation has been applied by the ERD, SemEval, and the 2014 version of the NEEL challenge.

On the opposite of an end-to-end evaluation is a step-by-step one. 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. without any error propagated from a previous step) and its output is evaluated. Despite the robustness of this approach, it is very time consuming, in particular, if participant systems use a different workflow than the typical one. One solution to this problem is to use a partial end-to-end evaluation.

A partial end-to-end evaluation aims to evaluate the output of each Entity Linking step by analysing the final result of the whole system. This evaluation uses different metrics that are influenced only by specific parts of the 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 evaluates only the correct mentions recognized, whereas another metric measures the performance of the NIL Clustering. This type of evaluation has been performed by TAC-KBP and, due to its benefits for the research community, has also been applied in the NEEL challenge since 2015.

Another aspect that differs in each challenge and in each version of the same challenge are the Entity Linking steps they require. The NEEL challenge was built by including a new step in each version. The 2013 version was purely based on Mention Recognition and Entity Typing. In 2014, the Candidate Selection step was included. In 2015 and 2016 the NIL Clustering was also proposed for tweets. The Candidate Detection step, however, has never been proposed for evaluation within the NEEL challenge because it is a hard step to evaluate from the final result of an Entity Linking system.

The TAC-KBP challenge also requires all the steps in the Entity Linking workflow, whereas the SemEval challenge and the long text track of the ERD challenge do not require the NIL Clustering step. The short text track of the ERD challenge requires only Mention Recognition and Candidate Detection.

For end-to-end evaluations, precision, recall, and F-measure are the evaluation metrics chosen to rank participants. In partial end-to-end, these metrics are also used for most of the steps. However, metrics to eval-

uate coreference chains have been proposed to evaluate the NIL Clustering and the conjunction of Mention Recognition and Candidate Selection. Two of these metrics proposed by TAC-KBP and further applied in NEEL were B^3 [44] and $CEAF_m$ [45]. A deeper explanation of these metrics is provided in Section 6.

3. 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. The ontology for annotating the entities changed from the flat CoNLL taxonomy consisting of 4 entity types (Person, Location, Organisation, Miscellaneous) to a much larger NERD ontology[15], which consists of 120 concepts grouped in two layers: the so-called core (containing most frequent classes) and the extended (containing sub-classes of the core classes).

The initial 2013 challenge dataset contains 4,265 tweets collected from the end of 2010 to the beginning of 2011, annotated according to the CoNLL taxonomy. These tweets cover a a variaty of topics, including comments on the news and politics. The dataset was split into 66% training and 33% test.

The second 2014 challenge dataset comprises 3,505 event-annotated tweets collected as part of the Redites project[16] from 15th July 2011 to 15th August 2011 (31 days), annotated according to the entire NERD ontology (core and extended), where each entity mention was linked to its corresponding DBpedia URI. The tweets were extracted from the Twitter firehose via a selection of hashtags. This dataset extends over multiple noteworthy events 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 third 2015 challenge dataset extends the 2014 dataset by typing the entity mentions with NERD ontology core classes and adding NIL references. This dataset consists of tweets published over a longer period, between 2011 and 2013. In addition to this, we 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 fourth 2016 challenge dataset builds on the 2014 and 2015 datasets, comprising 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.

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.

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 mentioning 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.

Summary statistics of the entity types are provided in Table 3, 4, and 5 respectively for the 2013, 2015, and 2016 corpora.[17] The most frequent entity type across all the datasets is Person. This is followed by Organisation and Location in the 2013 and 2015 datasets. In the

[15] http://nerd.eurecom.fr/ontology/nerd-v0.5.n3

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

[17] 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.

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	0.32	0.09

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.

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 3

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 4

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 5

Entity type statistics for the three data sets from 2016.

2016 dataset the second and third most frequent types are Product and Organisation. The distributional differences between the entity types in the three sets can be clearly seen. This makes the NEEL task challenging, particularly when tackled with supervised learning approaches.

The annotation of each tweet in the training and development datasets gave all participants a common base from which to learn extraction patterns. The test datasets contained no annotations; the challenge tasks were for participants to provide these. To assess the performance of the submissions we used an underlying gold standard. The gold standards were generated by multiple annotators. 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 linguists; 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 comprised the following steps:

Phase 1. The corpus was split into four quarters, each being annotated with 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 better quality and higher consensus in the annotations.

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.

The annotation process for 2014, 2015 datasets comprised 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 [37].

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, and GATE [38] for the 2015 challenge 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 till most problematic cases were resolved.

Phase 4. Particular to the 2015 challenge, unsupervised NIL Clustering generation, based on mention strings and their types, was performed.

Phase 5. Particular to the 2015 challenge, the third annotator went through all NILs to include or exclude them from a given cluster.

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 as the union of the 2014 and 2015 corpora. The training set was built on top of the 2014 and 2015 datasets, this meant to provide continuity with previous years and to build upon existing findings to push further the research. The 2016 challenge used the NEEL Challenge Annotation Guidelines provided in 2015. The test set was partially manually annotated (this accounts for 10% of the entire set).[18] The random selection was performed while preserving the original distributions of types in the corpus. The annotation process for the 2016 test set comprised 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 analyze 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 2. Consistency checking, the two annotators 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 iterated further Phase 1 untill most problematic cases were resolved.

Phase 3. Unsupervised NIL Clustering generation, based on mention strings and their types, was performed.

Phase 4. One of the two expert annotator went through all NILs to 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 made as required.

The lessons learned from building high quality gold standards are that the annotation process must be guided with Annotation Guidelines, at least two annotators must be involved in the annotation process to ensure consistency, and the feedback from the crowd

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

(participants) is valuable in improving the quality of the datasets, providing complementary annotations to the cases found by experts. The Annotation Guidelines, written by experts, must describe the annotation task (entity types and NERD ontology) 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 favoured GATE because the annotation process was guided by an ontology-centric view. The annotation task took less time having annotators with the same background.

4. Corpus Analysis

While the main goal 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 [70] 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, which we call CE2013, does not contain entity links.

4.1. Entity Overlap

Table 6 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.

4.2. Confusability

We define the true confusability of a surface form s as the number of meanings that this surface form can have.[19] 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 7. 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.

4.3. Dominance

We define the true dominance of a resource r_i 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 8. 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.

4.4. Readability

To gain an understanding of the difficulty of a text, several readability measures have been developed. In this subsection, we describe the most common measures. The scores for each on the NEEL corpora are presented in Table 9.

Flesch-Kincaid [67] Initially the Flesch-Kincaid measure was developed by the US Navy to estimate the difficulty of technical manuals. It is currently often used for official documents such as those in the law and insurance domain. Its score corresponds to a US school grade level and is computed as:

$$11.8 * \frac{syllables}{words} + 0.39 * \frac{words}{sentences} - 15.59 \quad (1)$$

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

	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 6

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 datasets pair the overlap is given in number of entities and percentage (in parentheses).

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

Table 7

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

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

Table 8

Dominance stats for analysed datasets.

Automated Readability Index (ARI) [69] The ARI index was also developed by the US military, and contrary to the Flesch-Kincaid test it compares characters to gauge the word length instead of syllables. The obtained scores correspond to US school grade levels. Decimal scores are rounded up. It is computed as:

$$4.71 * \frac{characters}{words} + 0.5 * \frac{words}{sentences} - 21.43 \quad (2)$$

Coleman-Liau [64] Similar to ARI, Coleman-Liau uses characters instead of syllables. It also roughly corresponds to US school grade levels. It is computed as:

$$5.88 * \frac{characters}{words} - 29.5 * \frac{sentences}{words} - 15.8 \quad (3)$$

Flesch Reading Ease [65] The Flesch Reading Ease score was developed by Rudolf Flesch in 1979. In this index, the scores lie between 0.00 and 100.0 where a higher score indicates an easier text to read. The formula is as follows:

$$206.835 - 84.6 * \frac{syllables}{words} - 1.015 * \frac{words}{sentences} \quad (4)$$

Fog Index [66] The FOG index was created by businessman Robert Gunning and discerns between the proportion of sentences with 'easy' and 'difficult' words. This difficulty is defined by the number of syllables a word has, although one could argue that long frequent words are less difficult than short infrequent words. It is computed using the following formula and its score also corresponds to the number of years of education deemed necessary to understand a text.

$$0.4 * \frac{words}{sentences} + 100 * \frac{words >= 3syllables}{words} \quad (5)$$

LIX [63] LIX was developed in Sweden. Rather than relying on the number of syllables or absolute character counts to distinguish long words, it computes the proportion of words that are over 6 characters and it is one of the few readability measures developed for languages other than English. The score can range between 20 and 60, with a higher score indicating that a text is more difficult. It is computed as:

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

SMOG grading [68] The SMOG reading formula was developed as a fix to the Fog index. It is computed as:

$$\sqrt{\frac{words >= 3syllables}{sentences} * 30} + 3 \quad (7)$$

Corpus	Flesch Kincaid	ARI	Coleman-Liau	Flesch Index	Fog Index	LIX	SMOG
CE 2013	4.1	4.5	7.2	84.8/100	6.4	32.4	7.1
NEEL 2014	5.9	6.4	7.6	79.8/100	8.9	32.7	8.6
NEEL 2015	6.0	6.4	7.5	79.7/100	9.0	32.6	8.7
NEEL 2016	6.0	6.7	8.6	76.4/100	8.9	33.7	8.9

Table 9
Readability scores for analysed datasets.

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

4.5. Summary

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

As the datasets are built on top of each other, they show a fair amount of overlap in entities between each other. This 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.

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

5. Approaches and Emerging Trends

Thirty different approaches were accepted in the four editions of the challenge starting from 2013. Table 10 lists all ranked teams. The approaches have proposed several differences, but we have observed some emerging trends that are uniquely to the top performing named entity recognition and linking approaches dealing with tweets. The main trend observed 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 showed a high dependence on the training data. This is not surprising, since these approaches are supervised, but this clearly defines that, to reaching top performance, labeled data is necessary. In addition, the large use of knowledge bases as dictionaries of entities featured by type has significantly affected the performance over the years. This strategy overcomes the lexical limitations of a tweet, but still preserves good results in the identification of entities available in the knowledge base used as referent. A common phase in all submitted approaches is normalisation, which is meant to normalise the lexical variations of the tweets and to translate them to language structures that can be better parsed by state of the art approaches. While the linguistic workflow favours the use of sequential solutions, the 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 novel 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.

From a historical perspective, starting from the first NEEL challenge on Concept Extraction (CE) until the 2016 edition, we observe:

- 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;

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

Table 10
Per year submissions and number of runs for each team.

- the use of highly performing end-to-end approaches for the candidate selection. Such a methodology has been further developed with the addition of fuzzy distance functions operating over ngrams and acronyms;
- the use of a pruning stage to filter out candidate entities. This has been presented in various approaches ranging from Learning-to-Rank to classification task. The latter showed best performance, holding more complexity in the definition of the feature sets;
- the use of hierarchical clustering of mentions meant 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, largely used in the first editions of NEEL, but later on this has shown to have limited performance as the task became more constrained.

Table 11 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, we add Entity Typing that has been jointly proposed as component of Mention Detection and Candidate Selection.

Challenge tasks incrementally changed over the years. The first challenge task in 2013 focused on Mention Detection combined with Typing. This task was then extended in 2014 to Mention Detection combined with Candidate Selection and the consecutive years added more levels of complexity to the task, leading to the joint analysis over the Mention Detection and Linking on both in-knowledge base and external entities. In the remainder of this analysis we will focus on two main tasks, namely Mention Detection and Candidate Selection. Table 12 presents a detailed description of the approaches taken for the Mention Detection combined with Typing. Participants approached the task using rule-based, machine learning and hybrid methods. For 2013, the strategies yielding the best results where hybrid, where models relied on the application of off-the-shelf systems (e.g., AIDA [46], ANNIE [47], OpenNLP,[24] Illinois NET [48], Illinois Wikifier [49], LingPipe,[25] OpenCalais, Stanford NER [50], WikiMiner,[26] NERD [51], TwitterNLP [53], AlchemyAPI, DBpedia Spotlight, Zemanta) for both the identification of the boundaries of the entity (mention detection) or the assignment of a semantic type (entity typing). The top performing system resulted to be System 1. For 2014 the strategy building the best result (System 14) was rule-based. In terms of features, the use of ngrams, gazetteers and part-of-speech (POS) features played an important role in the system's

[24] https://opennlp.apache.org
[25] http://alias-i.com/lingpipe
[26] 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, capitalization, 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, 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[20]
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,[21] OpenCalais,[22] Zemanta[23]
Candidate Selection	Distributional Semantic Model, Random Forest, RankSVM, Random Walk with Restart, Learning to Rank	gloss, contextual features, graph distance	Wikipedia, DBpedia	DBpedia Spotlight [52], AlchemyAPI, Zemanta, Babelfy [55]
NIL Clustering	Conditional Random Fields, Random Forest, Brown Clustering, Lack of candidate, Score Threshold, Surface Form Aggregation, Type Aggregation	POS, contextual words, n-grams length, predicted entity types, capitalization ratio, entity mention label, entity mention type		

Table 11

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

performance. The 2015 best performing approach for Mention Detection was largely inspired by the 2014 winning approach: the use of ngrams used to look up resources in DBpedia and a set of lexical features such as POS, position in tweets. The type was assigned as output of a classification task over the sub-set of tokens being mapped to DBpedia resources. We can observe how the complexity of the task has been postponed to the next stage of the workflow, namely Candidate Selection. This is to favor the recall of the mention identification phase, and thus having larger set of candidates for the final linking phrase. The 2016 best performing system, System 26, implements a pure end-to-end solution where all surface forms of the entities in the knowledge base are used to look up the unigram taken from the tweets.

Following the 2013 challenge, the task for the following year was enhanced by asking participants to link the extracted entities to their corresponding DBpedia link (if exists). Table 13 describes the approaches taken by the 2014, 2015, 2016 participants. For the 2014 challenge, most of the candidates approached the Candidate Selection task sequentially. However, the best performing system (System 14) approached it as a joint task and proposed the so-called end-to-end. As opposed to most of the participants which used off-the-shelf tools, System 14 provided a from-scratch rule-based approach using a combination of machine learning gradient boosting approaches. Some systems ap-

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	Rule-based	Wiki and DBpedia Gazetteers
8	NLTK [54]	ngrams, NLTKPos	Rule-based	Wikipedia
9	Babelfy API [55]	-	Rule-based	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	Rule-based	Wikipedia and Freebase lexicons
15	TwiNER [58]	Regular Expression, Entity phrases, N-gram	TwiNER and CRF	DBpedia Gazetteer, Wikipedia
16	TAGME [56]	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 [57]
	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 [59]	-	-	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 [60]	tokens and POS	CRF	-
30	-	tokens	Lexical Similarity and Rule-based	-

Table 12

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

plied 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 for the task there are: NER dictionaries (e.g., Google Cross-Wiki); Knowledge Base Gazetteers (e.g., Yago, Dbpedia); weighted lexicons (e.g., Freebase, Wikipedia); other sources (e.g., Microsoft Web N-gram).[27] A wide range of features were investigated for the 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. For the 2015 challenge, System 20 (ranked first) proposed an enhanced version of the end-to-end system winner of the 2014 edition, 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 Fields (CRF) classifier. The other approaches can be classified as sequential, where the complexity is moved to only performing the right indexing of the entity in the knowledge base. Most of these approaches exploit the popularity of the entities and apply distance similarity functions to better rank entities. From the analysis the move to pipeline controlled in-house solutions 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 have exploited lexical similarity distance functions and class information of the mention. In 2016, the winning System 26, proposed a joint Mention Extraction and Candidate Selection, where ngrams of the text are mapped to DBpedia entities. A preprocessing stage cleans and normalizes the initial tweets, and scoring measures, weighting graph distance measurements, connected component analysis, centrality of the entities and density observations are used to resolve the selection of entities in case of ambiguity. Such an approach adheres to the before winning approaches, but it does not implement a pruning stage. System 27, ranked second, implements a linguistic pipeline, where the Candidate Selection is performed by looking up entities according to the exact lexical value of the mentions with DBpedia titles, redirect pages, and disambiguation pages. The main part of the approach is the preprocessing that consists of normalizing the input text and making it similar to formal language text. The other trend we can observe is that only System 28 exploited an external system for the Candidate Selection, while all others implement in-house solutions.

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

6. Evaluation Strategies

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

6.1. 2013 Evaluation Measures

In 2013, the submitted systems were evaluated based on performance in extracting a mention and assigning its correct class from a test set TS. For each instance in TS a system was requested to provide a set of tuples of the form: $T_i = (m, t)$, where *m* is the mention and *t* is the type, which were then compared against tuples in the gold standard (GS). A type is any valid materialization of the class defined in the challenge guidelines and defined as Person-type, Organization-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.

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; $(m, t) \in GS_T$ 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:

$$TP_T = \{(m,t)_S | (m,t)_{GS} \in (S_T \cap GS_T)\} \quad (8)$$

$$FP_T = \{(m,t)_S | (m,t)_{GS} \in S_T \wedge (m,t) \notin GS_T\} \quad (9)$$

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[62] and MART gradient boosting	Wikipedia, Freebase
15	Google Search	ngrams, DBpedia and Wikipedia links, capitalization	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, normalized 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 [61]	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 13
Presents per year submissions and number of runs for each team for the Candidate Selection phase.

$$FN_T = \{(m,t)_S | (m,t)_{GS} \in GS_T \wedge (m,t) \notin S_T\} \tag{10}$$

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:

$$P_T = \frac{|TP_T|}{TP_T \cup FP_T} \tag{11}$$

$$R_T = \frac{|TP_T|}{TP_T \cup FN_T} \tag{12}$$

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

$$P_T = \frac{P_{PER} + P_{ORG} + P_{LOC} + P_{MISC}}{4} \tag{13}$$

$$R_T = \frac{R_{PER} + R_{ORG} + R_{LOC} + R_{MISC}}{4} \quad (14)$$

$$F_1 = 2 \times \frac{P_T \times R_T}{P_T + R_T} \quad (15)$$

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

6.2. 2014 Evaluation Measures

In 2014, a system S was evaluated in terms of its performance in extracting both mentions and links from tweets of the test set (TS). For each instance in TS, a system provided a tuple of the form: mention (m), and link (l). A link is any valid DBpedia URI[28] 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 the gold standard 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.

The evaluation performed *a prior* normalization stage over each submission, resolving where needed, the redirects. Then, it assessed the correctness of tuples provided by a system S as the exact-match of the mention and the link. Tuple order was also relevant. 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:

$$TP_T = \{(m,l)_{TS} | (m,l)_{GS} \in (S_T \cap GS_T)\} \quad (16)$$

$$FP_T = \{(m,l)_{TS} | (m,l)_{GS} \wedge (S_T \cap GS_T)\} \quad (17)$$

$$TP_T = \{(m,l)_{TS} | (m,l)_{GS} \wedge (S_T \cap S_T)\} \quad (18)$$

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

Thus TP defines the set of relevant pairs in TS, in other words the set of pairs in TS that match corresponding ones in GS. FP is the set of irrelevant pairs in TS, in other words the pairs in TS that do not match the pairs in GS. FN is the set of false negatives denoting the pairs that are not recognised by TS, yet appear in GS. Since 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 entity e recognised to the correct resource *l*. Precision, Recall, F_1 were defined as in Eq. 11, Eq. 12, Eq. 15 respectively.

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

6.3. 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 19) that, in 2015, also contained the computation time of an entry in case two systems had exactly the same performance.

$$\begin{aligned} score = {} & 0.4 * mention\_ceaf \\ & + 0.3 * strong\_typed\_mention\_match \\ & + 0.3 * strong\_link\_match \end{aligned} \quad (19)$$

The *strong_typed_mention_match* measures the performance of the system regarding the correct identification of mentions and their correct type assertion. The detection of mentions is still based on strict matching as in previous versions of the challenge. Therefore true positive (Equation 8), false positive (Equation 9), and false negative (Equation 10) 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, the more mentions in a tweet the less a single mention impact the result. In 2015 we used micro-averaged metrics. In micro-averaged precision and recall each mention has an equal impact on the final result, regardless of how many mentions appear in the same tweet. Therefore, precision (P_{TM}) is calculated according to the Equation 20 and recall (R_{TM}) according to Equation 21. Finally, *strong_typed_mention_match* is the micro-averaged (F_1) as given by Equation 22.

$$P_{TM} = \frac{\sum_t |TP_t|}{\sum_t TP_t \cup FP_t} \tag{20}$$

$$R_{TM} = \frac{\sum_t |TP_t|}{\sum_t TP_t \cup FN_t} \tag{21}$$

$$strong\_typed\_mention\_match = 2 \times \frac{P_{TM} \times R_{TM}}{P_{TM} + R_{TM}} \tag{22}$$

The *strong_link_match* metric measures the correct link between a correctly recognized mention and a knowledge base entry. This metric is also based on micro-averaged precision and recall, however the true positive, false positive, and false negative also includes the link step. Let $(x, y) \in S_l$ denote the set of tuples corresponding to the link l provided by system S; $(x, y) \in GS_l$ denote the set of tuples regarding the links in the gold standard. Then the set of true positives (TP_l), false positives (FP_l), and false negatives (FN_l) regarding the linking between mentions and Knowledge Base entries are given by Equations 23, 24, and 25.

$$TP_l = \{(x, y) | (x, y) \in (S_l \cap GS_l)\} \tag{23}$$

$$FP_l = \{(x, y) | (x, y) \in S_l \wedge (x, y) \notin GS_l\} \tag{24}$$

$$FN_l = \{(x, y) | (x, y) \in GS_l \wedge (x, y) \notin S_l\} \tag{25}$$

To be considered a correct link a system must have detected a mention and its type correctly (x) and the correct Knowledge Base entry (y). Note also that this metric does not evaluate links to NIL. Precision and recall for the linking step are then calculated as expressed in Equations 26 and 27. Finally, *strong_link_match* is given as in Equation 28.

$$P_L = \frac{\sum_l |TP_l|}{\sum_l TP_l \cup FP_l} \tag{26}$$

$$R_L = \frac{\sum_l |TP_l|}{\sum_l TP_l \cup FN_l} \tag{27}$$

$$strong\_link\_match = 2 \times \frac{P_L \times R_L}{P_L + R_L} \tag{28}$$

The last metric in our evaluation score is given by the Constrained Entity-Alignment F-measure (CEAF) [45]. This is a metric that measures coreference chains and is used to jointly evaluate Candidate Selection and NIL Clustering steps. Let $E = \{m_1, \ldots m_n\}$ 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, in order to submit (up to 10) contending entries to a registry of the NEEL challenge services. In this context, we refer to a contending entry as the participant's REST endpoint queried in the evaluation campaign. Each endpoint had a Web address (URI) and a name, which was referred as runID. Upon receiving the registration of the REST endpoint, calls to the contending entry were scheduled in two different time windows, namely, D-Time - to test the APIs, and T-Time - for the final evaluation and metric computations. To ensure correctness of the results and avoid any loss, we triggered a large number of queries and statistically evaluated the results. Details

of the algorithm is provided in [3]. 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 19.

In 2016, submissions were evaluated offline as participants were asked to annotate the TS during a short time window and to send the results in a TSV file.

Three editions out of four followed a conventional offline evaluation process. 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 goodness 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 tweet.

7. Conclusion

The NEEL series of challenges is a research endeavour established in 2013 to foster the development of novel automated approaches for mining semantics from tweets while ensuring a convergence of the community towards a standardized set of benchmark corpora, thus offering comparable means to practitioners.

The rigorous procedure used to create the corpora has offered to a large community high quality labeled datasets that can be used as inputs of automated supervised approaches. The procedure has been incrementally iterated and adjusted over the four years, providing continuity with the past and thus ensuring the re-usability of the approaches over the different editions. While the consolidation has provided consistent labeled data, it has also showed the robustness of the community being built and becoming a gold reference in the field. To the best of our knowledge these are the largest publicly available corpora providing named entities, types, and link annotations for tweets. To maximise the impact, we have released the corpus yearly via sharing the annotations and the tweet IDs with public license.

Despite the rigorous human annotation protocol, the datasets have a low variance of semantics in the entity definition since there are few entities in our datasets with many different referents. We have conducted preliminary studies on readability of the tweets to quantify the complexity of the annotation task, but we obtained partial results not particularly representative for this domain and thus it points out future research activities.

Participants proposed distinctive submissions, aiming to strengthen the peculiar know-how of the team. Nevertheless, we have witnessed a convergence of the approaches towards data-driven solutions where knowledge bases are prominently used to discover known entities and labeled data used to select the candidates and suggest novel entities. Such an approach, first proposed in 2014 by System 14 has become, with variations, the leading solution also for the following years. Despite the consolidated number of options for addressing the challenge task, the participants' results show that the NEEL task remains challenging when applied to tweets with their peculiarities, compared to standard, lengthy texts and thus more work will be done.

With the ever spreading initiatives on this matter, the NEEL series of challenges has adapted the evaluation strategies to ensuring fairness of the evaluation, transparency, and correctness. This has triggered the use of scoring tools in 2013 and 2014. These tools were made publicly available and discussed in communities, while from the 2015 the use of the scorer adopted in the TAC-KBP challenge. In 2015, we have extended the scorer to take into account the computing time. This experiment led to interesting observations but moved the complexity of the task to the distributed computing and optimisation that go beyond the typical know-how of practitioners in the field. This suggested to propose a conventional evaluation strategy the year after while ensuring the fairness of the results asking to produce large scale annotations in a short time window.

Beyond the thirty teams who completed the evaluations in four years, more than three hundred participants got in contact with the NEEL organisers to explicitly acquire the corpora according to the running yearly schedule. The teams come from more than twenty different countries and are both from academia and industry. In fact, the 2014 and 2015 winners are companies operating in the field, respectively Microsoft and Studio Ousia. The 2013 and 2016 winners

are academic teams. The outreach of the NEEL series of challenge is also witnessed by the grants offered by companies (ebay[29] in 2013 and SpazioDati[30] in 2015) and research projects (LinkedTV[31] in 2014, and FREME[32] in 2016).

The NEEL series of challenges also triggered the interest of localised communities such as the NEEL-IT[33] that is deploying the NEEL guidelines (with minor variations due to the intra-language dependencies) and know-how to establish a community and create a standardised benchmark for sharing the algorithms and results of mining semantics from Italian tweets. This is a first step toward a multilingual NEEL challenge. In 2015, we also built bridges with the TAC community. We plan to strengthen these and to involve a larger audience of potential participants spanning the Linguistics, Machine Learning, Knowledge Extraction and Data and Web Science fields, in order to widen the scope for potential solutions to what is acknowledged to be a challenging and valuable exercise.

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, *NEEL'16: Named Entity rEcognition & Linking Challenge Report*, 6^{th} International Workshop on Making Sense of Microposts, 2016.

[5] 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.

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

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[29] http://www.ebay.com
[30] http://www.spaziodati.eu
[31] http://www.linkedtv.eu
[32] http://freme-project.eu/
[33] http://neel-it.github.io

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

[35] R. C. Bunescu, M. Pasca, Using Encyclopedic Knowledge for Named entity Disambiguation, EACL, 2006.

[36] S. Cucerzan, Large-Scale Named Entity Disambiguation Based on Wikipedia Data, EMNLP-CoNLL, 2007.

[37] 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.

[38] 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.

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

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

[41] D. Carmel, M.-W. Chang, E. Gabrilovich, B.-J. P. Hsu, K. Wang, *ERD'14: Entity Recognition and Disambiguation Challenge*, ACM SIGIR Forum, 2014.

[42] A. Moro, R. Navigli, *SemEval-2015 task 13: multilingual all-words sense disambiguation and entity linking*, Proceedings of SemEval, 2015.

[43] 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*, Proceedings of the the 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.

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

[45] X. Luo, *On coreference resolution performance metrics*, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP), 2005.

[46] 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*, Conference on Empirical Methods in Natural Language Processing, 2011.

[47] 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.

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

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

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

[56] 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.

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

[58] 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 ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2012.

[59] 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.

[60] 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.

[61] 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.

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

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

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

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

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

[67] 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.

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

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

[70] 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.